NTCIR-12 MathIR Task Overview

Richard Zanibbi
Rochester Institute of Technology
rlaz@cs.rit.edu

Akiko Aizawa
National Institute of Informatics
aizawa@nii.ac.jp

Michael Kohlhase
Jacobs University Bremen
m.kohlhase@jacobs-university.de

Iadh Ounis
University of Glasgow
Iadh.Ounis@glasgow.ac.uk

Goran Topić
National Institute of Informatics
goran_topic@nii.ac.jp

Kenny Davila
Rochester Institute of Technology
kxd7282@rit.edu

ABSTRACT

We present an overview of the NTCIR-12 MathIR Task, dedicated to information access for mathematical content. The MathIR task makes use of two corpora. The first corpus contains excerpts from technical articles in the arXiv, while the second corpus contains English Wikipedia articles. For each corpus, there were two subtasks. Three subtasks contain queries with keywords and formulae (*arXiv-main*, *Wiki-main*, and *arXiv-simto*), while the fourth considers isolated formula queries (*Wiki-formula*). In this overview paper, we summarize the task design, corpora, submitted runs, results, and the approaches used by participating groups.

Subtasks

MathIR arXiv Main Task (English), optional MathIR arXiv Similarity Task (English), optional MathIR Wikipedia Task (English), optional MathIR Wikipedia Formula Browsing Task (English)

Keywords

Mathematical Information Retrieval (MIR), MathML, Query-by-Expression

1. INTRODUCTION

This task aims to support research in Mathematical Information Retrieval (MIR) and its related fields [5,17]. Mathematical formulae are important means for dissemination and communication of scientific information. They are used for both calculation and clarifying definitions and explanations given in natural language. Despite the importance of math in technical documents, most search engines do not support users' access to mathematical formulae in target documents.

This paper summarizes the third math retrieval task at NTCIR. The NTCIR-10 Math Pilot Task [1] was a first attempt to develop a common workbench for mathematical formula search. For the subsequent NTCIR-11 Math-2 task [2], we continued to pursue our initial goal of creating a shared evaluation platform for an active and emerging community in Math IR. This was a traditional ad-hoc retrieval task with formula + keyword queries. As a subtask of Math-2, the Wikipedia subtask provided the first forum for comparing formula search engines, based upon their ability to retrieve specific formula in documents [12]).

For the NTCIR-12 MathIR task, we have created a new corpus of Wikipedia articles containing mathematical formula, along with four new search tasks. We created the new corpus to provide mathematical information useable by non-experts, and to explore search topics for this large and important user group. An experimental task was developed to test a new formula query operator, the *simto region*. We also created new topics for the NTCIR-11 arXiv corpus comprised of small excerpts from technical articles. Finally, our new formula search task considers relevance assessments for formula search rather than recall for specific targets as used in the NTCIR-11 Wikipedia formula subtask.

In the remainder of this paper we summarize the MathIR task design (Section 2), participant systems (Section 3), and present and discuss task results (Section 4).

2. TASK DESIGN

2.1 Corpora

Two corpora were used for the MathIR task. The first contains paragraphs from technical articles in the arXiv,[1] while the second contains complete articles from Wikipedia. Generally speaking, the arXiv articles are written by technical experts assuming some level of mathematical sophistication from readers. In contrast, many Wikipedia articles on mathematics are written to be accessible for non-experts.

arXiv Corpus. The arXiv dataset for NTCIR-12 MathIR is the same one used for the NTCIR-11 Math-2 task [2]. It consists of 105,120 scientific articles in English. These articles were converted from LaTeX to an HTML+MathML-based format by the KWARC project.[2] The dataset contains articles from the arXiv categories math, cs, physics:math-ph, stat, physics:hep-th, physics:nlin to get a varied sample of technical documents containing mathematics.

Each document is divided into paragraphs, and we use these as the return units ('documents') for the task. This produces 8,301,578 search units with roughly 60 million math formulae (including isolated symbols). Excerpts are stored independently in separate files, in both HTML5 and XHTML5 formats. Additional information is available elsewhere [2].

Wikipedia Corpus. This new corpus was created for the MathIR task. The MathIR Wikipedia corpus contains 319,689 articles from English Wikipedia converted into a simpler XHTML format with images removed (5.15 GB un-

[1] http://www.arxiv.org
[2] http://kwarc.info/

compressed).[3] Unlike the arXiv corpus, articles are not split into smaller documents. 10% of the sampled articles contain explicit `<math>` tags that demarcate LaTeX. All articles with a `<math>` tag are included in the corpus. The remaining 90% of the articles are sampled from Wikipedia articles that do not contain a `math` tag. These 'text' articles act as distractors for keyword matching, and reflect the small proportion of articles related to math in Wikipedia, while keeping the corpus size manageable for participants.

There are over 590,000 formulae in the corpus, encoded using LaTeX, Presentation MathML and Content MathML. Formulae were encoded using a pipeline similar to that used to construct the arXiv corpus, with an additional step to convert mediawiki templates for mathematics to LaTeX. Note that untagged formulae frequently appear directly in HTML text (e.g. `'where x <sup> 2 ...'`). We made no attempt to detect or label these formulae embedded in the main text.

2.2 Topics

For the NTCIR-12 Math-12 subtasks, we generated 107 search topics and distributed the set to the participants in a custom XML format. A summary of the topics for each subtask is shown in Tables 1, 2 and 3. Along with the number of keywords and formulae in each query, these tables provide the number of nodes and maximum depth in MathML tree representations for formulae, along with the number of query variables (wildcards) and *simto* regions (see below) included in query formulae.

Topic Format. For participants, a MathIR topic contains: (1) a Topic ID, and (2) a Query (formula + key words), but no textual description. The description is omitted to avoid participants biasing their system design towards the specific information needs identified in the topics. For evaluators, each topic also contains a narrative field that describes a user situation, the user's information need, and relevancy criteria. Formula queries are encoded in LaTeX, Presentation MathML, and Content MathML. Further details about the topic format are available elsewhere [6].

arXiv Topics. Many of the topics in the arXiv-main task are sophisticated, for example seeking to determine whether a connection exists between a factorial product and products starting with one (MathIR-2). Some queries are simpler, such as looking for applications of operators, or loss functions used in machine learning. This task has 29 topics.

Queries in the arXiv Similarity Task (Table 2) combine keywords with formulae containing an operator identifying subexpressions that may be 'similar to' rather than identical to the query. As with the arXiv-main task, these queries are designed with mathematically sophisticated users in mind. This task was experimental, and contains 8 topics.

Wikipedia Topics. Topics for the Wikipedia main task have been designed with a less expert user population in mind. We imagined undergraduate and graduate students searching Wikipedia to locate or relocate specific articles (i.e. navigational queries), browse math articles, learn/review mathematical concepts and notation they come across in their studies, find applications of concepts, or find information to help solve particular mathematical problems (e.g., for homework). There are 30 topics for this task. The narrative scenarios detailing how to assess relevance all state that articles linking to a relevant article are considered to be *partially* relevant.

For the Wikipedia Formula Browsing task, we consider users browsing formulae using isolated formulae as queries. Relevant formulas are those felt to be be similar in appearance and/or mathematical content to the query formula. There are no keywords in the Wiki-formula task (see Table 3). There are 40 formula queries in total: the first 20 queries are *concrete* without wildcards, and the remaining 20 queries contain wildcards. Queries 21-40 are produced from the first 20 queries by deleting subexpressions and/or replacing subexpressions with wildcards.

Formulae Query Variables (Wildcards). Formulae may contain query variables that act as wildcards, which can be matched to arbitrary subexpressions on candidate formulae. Query variables were represented using two different representations for the arXiv and Wikipedia topics. For the arXiv tasks, query variables are named and indicated by a question mark (e.g., *?v*) while in Wikipedia wildcards are numbered and appear between asterisks (e.g., **1**).

Here is an example query formula with three query variables, ?f, ?v, and ?d.

$$\frac{\mathsf{?f}(\mathsf{?v}+\mathsf{?d})-\mathsf{?f}(\mathsf{?v})}{\mathsf{?d}} \tag{1}$$

This query matches the argument of the limit on the right-hand side of the equation below, substituting g for ?f, cx for ?v, and h for ?d. Note that each repetition of a query variable matches the same subexpression.

$$g'(cx)=\lim_{h\to 0}\frac{g(cx+h)-g(cx)}{h} \tag{2}$$

Formula Simto Regions. *Similarity regions* modify our formula query language, distinguishing subexpressions that should be identical to the query from those that are similar to the query in some sense. Consider the query formula below, which contains a *similarity region* named 'a.'

$$\frac{\boxed{\text{a}}\ \overline{g(cx+h)-g(cx)}}{h} \tag{3}$$

Here the fraction operator and h should be matched exactly, while the numerator may be replaced by a 'similar' subexpression. Depending on the notion of similarity we choose to adopt, simto region 'a' might match '$g(cx+h)+g(cx)$', if addition is similar to subtraction, or '$g(cx+h)-g(\mathbf{d}x)$', if c is somehow similar to d. Simto regions may also contain exact match constraints (see [6]).

2.3 Participant Submissions

Given a query, participant systems estimate the relevance of 'documents' in the corpus to the query (paragraphs for arXiv tasks, articles for Wikipedia tasks), and then return a ranked list of documents. For each task, participants could submit up to four runs with 1,000 results per query. Results include the score for each returned document along with supporting evidence (e.g. the formula identifier, keywords, or substitution terms for query variables and simto regions). Hit justifications are used to assist the evaluators, for example by highlighting specified formula regions and keywords in the evaluation interface (see Figure 1). Submissions were provided in a custom XML format [6], which was later converted into a standard `trec_eval` format by the organizers. To assist with result reporting, a submission validation script was distributed to the participants.

[3] `http://www.cs.rit.edu/~rlaz/NTCIR12_MathIR_WikiCorpus_v2.1.0.tar.bz2`

Table 1: Topics for Main Tasks (Keywords + Formulae). *PMML Nodes* represents the number of nodes for all query formulae in Presentation MathML. *Max Depth* is the maximum depth of an expression tree in PMML. *Num qvar* is the number of wildcards (query variables) in formulae.

arXiv Main Task ('Main')

Topic ID	Num keywords	Num formulae	PMML nodes	Max depth	Num qvar
MathIR-1	2	1	5	2	1
MathIR-2	2	1	4	2	1
MathIR-3	2	1	4	2	1
MathIR-4	2	1	4	2	1
MathIR-5	2	1	5	2	1
MathIR-6	2	1	18	4	4
MathIR-7	2	1	17	4	4
MathIR-8	2	1	6	3	1
MathIR-9	2	1	24	7	6
MathIR-10	2	1	18	4	3
MathIR-11	1	1	8	2	0
MathIR-12	0	1	28	4	7
MathIR-13	0	1	16	4	5
MathIR-14	1	1	7	3	0
MathIR-15	2	1	7	3	1
MathIR-16	0	1	21	5	3
MathIR-17	3	1	1	1	0
MathIR-18	1	1	28	5	1
MathIR-19	2	2	14	3	0
MathIR-20	2	2	28	4	3
MathIR-21	1	1	12	4	3
MathIR-22	1	1	9	4	2
MathIR-23	3	1	1	1	0
MathIR-24	3	1	13	4	2
MathIR-25	3	1	32	6	1
MathIR-26	3	1	24	8	7
MathIR-27	0	1	19	4	1
MathIR-28	1	1	9	2	0
MathIR-29	0	1	14	3	2

Wikipedia Task ('Wiki-main')

Topic ID	Num keywords	Num formulae	PMML nodes	Max depth	Num qvar
MathWiki-1	3	1	2	1	0
MathWiki-2	2	2	6	2	0
MathWiki-3	1	1	4	2	0
MathWiki-4	1	1	14	5	0
MathWiki-5	2	1	18	3	0
MathWiki-6	2	1	25	6	0
MathWiki-7	0	1	7	3	1
MathWiki-8	1	1	6	3	1
MathWiki-9	2	1	22	6	3
MathWiki-10	0	1	18	6	0
MathWiki-11	0	1	12	3	0
MathWiki-12	1	1	7	4	0
MathWiki-13	1	1	23	5	2
MathWiki-14	2	2	17	3	2
MathWiki-15	3	1	11	6	0
MathWiki-16	2	2	28	6	0
MathWiki-17	3	1	26	5	0
MathWiki-18	2	1	33	8	0
MathWiki-19	2	1	22	2	0
MathWiki-20	5	2	27	5	5
MathWiki-21	3	2	7	2	2
MathWiki-22	2	1	15	5	0
MathWiki-23	2	1	6	3	0
MathWiki-24	3	1	13	5	1
MathWiki-25	1	1	25	6	0
MathWiki-26	1	1	13	4	0
MathWiki-27	1	1	13	4	3
MathWiki-28	3	4	77	7	0
MathWiki-29	4	1	21	6	0
MathWiki-30	2	2	42	6	5

Table 2: Topics for arXiv Similarity Task.

Topic ID	Num keyw.	Num form.	PMML nodes	Max depth	Num qvar/simto
MathIR-1	0	1	16	5	2/1
MathIR-2	2	2	60	5	0/5
MathIR-3	2	1	13	5	3/1
MathIR-4	3	1	16	6	3/2
MathIR-5	4	1	32	5	6/1
MathIR-6	4	1	24	7	0/3
MathIR-7	2	1	45	8	6/1
MathIR-8	2	1	54	9	6/1

2.4 Evaluation Protocol

The evaluation of the MathIR task was pooling-based. First, all submitted results were converted into a `trec_eval` result file format. Next, for each topic, the top-20 ranked documents were selected from each run. Then, the set of pooled hits were evaluated by human assessors.

Evaluators. For the arXiv tasks, to ensure sufficient familiarity with mathematical documents, three evaluators were chosen from third-year and graduate students of (pure) mathematics. For the Wikipedia tasks, intended to represent mathematical information needs for non-experts, ten students were recruited for evaluation: five undergraduates, and five graduate (MSc) students. Each hit was evaluated by one undergraduate and one graduate student. To reduce bias, evaluations were rotated so that each undergraduate evaluated at least one hit with each graduate student and vice versa. This led to each student evaluating just two hits with the same student in the Wiki-main task, and two hits with three different students in the Wiki-formula task (due to the larger number of topics). All evaluators were briefly acquainted with the Sepia interface, and the query language prior to evaluating hits.

Evaluation Interface. After the pooling process, the selected retrieval units were fed into the SEPIA system [13] with MathML extensions developed by the organizers. Fig. 1 is a screenshot of the actual SEPIA system used for evaluation. The light red box at the top of the interface contains information on the topic, including query keywords and formulae, the title of the topic, a scenario description defining relevance assessment criteria, and an example hit link (if provided) is displayed. The lower-right white box shows the current document being evaluated along with the URL for the original arXiv article or (live) Wikipedia page.

Evaluators judged the relevance of each hit by comparing it to the query formulae and keywords, along with the described scenario provided with the topic. Relevance is evaluated for retrieved documents in the arXiv-main, arXiv-simto, and Wiki-main subtasks, and for individual formulae in the Wiki-formula subtask. To assist evaluators with determining the relevance of each document to the query, the keywords and formulae included as justifications in the submission files were highlighted on screen, as illustrated in Figure 1. For the Wiki-formula task, each returned formula was represented in a separate document, with the formula highlighted within the article where it appears.

Relevance Ratings. For each retrieval unit, the evaluators were asked to select either `relevant` (R), `partially-relevant` (PR), or `not-relevant` (N), using buttons located at the bottom of the document in the Sepia interface. Each

Table 3: Topics for Wikipedia Formula Browsing.

CONCRETE QUERIES

TOPIC ID	PMML NODES	MAX DEPTH
MathWikiFormula-1	5	3
MathWikiFormula-2	1	1
MathWikiFormula-3	20	7
MathWikiFormula-4	39	8
MathWikiFormula-5	28	14
MathWikiFormula-6	23	4
MathWikiFormula-7	24	4
MathWikiFormula-8	42	10
MathWikiFormula-9	53	8
MathWikiFormula-10	44	8
MathWikiFormula-11	15	5
MathWikiFormula-12	12	5
MathWikiFormula-13	19	6
MathWikiFormula-14	36	7
MathWikiFormula-15	30	4
MathWikiFormula-16	50	10
MathWikiFormula-17	29	4
MathWikiFormula-18	50	9
MathWikiFormula-19	239	10
MathWikiFormula-20	136	12

WILDCARD QUERIES

TOPIC ID	PMML NODES	MAX DEPTH	NUM QVAR
MathWikiFormula-21	6	3	1
MathWikiFormula-22	3	2	1
MathWikiFormula-23	14	6	1
MathWikiFormula-24	5	3	2
MathWikiFormula-25	13	7	2
MathWikiFormula-26	11	4	2
MathWikiFormula-27	12	4	3
MathWikiFormula-28	39	10	3
MathWikiFormula-29	28	7	7
MathWikiFormula-30	42	7	9
MathWikiFormula-31	15	5	3
MathWikiFormula-32	11	5	2
MathWikiFormula-33	12	4	3
MathWikiFormula-34	20	5	6
MathWikiFormula-35	15	3	2
MathWikiFormula-36	28	7	6
MathWikiFormula-37	20	3	5
MathWikiFormula-38	33	7	4
MathWikiFormula-39	19	9	3
MathWikiFormula-40	79	10	8

retrieval unit was rated by two assessors. Evaluators had to rely on their mathematical intuition, the described information need, and the query itself to determine hit ratings.

Since the `trec_eval` tool only accepts binary relevance judgments, the scores of evaluators were first converted into a combined relevance score using the mapping shown in Table 6. If the final rating is equal-or greater than three, the overall judgment is considered to be `relevant`; if greater than or equal to one, `partially-relevant`. When there were more than two hit ratings, we took the average and doubled it (59 out of 11,640 hits, 0.51%).

Evaluation Metrics. Precision@k for $k = \{5, 10, 15, 20\}$ was used to evaluate participant systems (see [16]). We chose these measures because they are simple to understand, and characterize retrieval behavior as the number of hits increase. Precision@k values were obtained from `trec_eval` version 9.0, in which they were labeled as `P_avgjg_5`, `P_avgjg_10`, `P_avgjg_15` and `P_avgjg_20`, respectively.

3. PARTICIPANT SYSTEMS

In this section, we briefly summarize the approaches used by task participants. As seen in Table 4, six groups submitted a total of 47 runs to the NTCIR-12 MathIR Task.All six participating teams submitted runs to the Wiki-main subtask. Systems were automatic except for the FSE team, which submitted one manual run where queries are manually edited, and hits are selected manually.

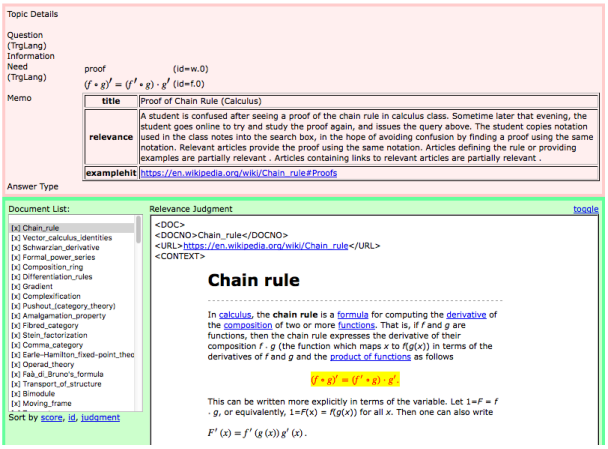


Figure 1: SEPIA Evaluation Interface Screenshot. Note the formula highlighted in yellow, identified by one of the participant systems as a 'hint.' Pooled hits for the query appear in a list at left.

Table 5 summarizes the configuration of participating systems. All participating systems used both keywords and formulae provided in queries. All provided formula encodings were used for the task. Whether the `tree structure` of formulae was used, and whether `query variables` were supported for math formulae varied by system. One group did not directly consider structural information for formulae, while another did not support query variables. All groups also used general-purpose search engines, with one group using a general-purpose search engine only for text retrieval.

In terms of architecture, most participant systems employ a re-ranking step, wherein one or more initial rankings are merged and/or re-ordered. This approach of obtaining an initial candidate ranking followed by a refined ranking is a common but effective strategy. To locate strong partial matches, all of the automated systems employ unification, whether for variables (e.g., '$x^2 + y^2 = z^2$' unifies with '$a^2 + b^2 = c^2$' [7]), constants [3], or entire subexpressions (e.g., via structural unification [10] or indirectly through *generalized terms* with wildcards for operator arguments [4, 15]).

The system descriptions in the remainder of this Section were contributed by the participating groups.

ICST (Peking University [4])

The ICST system is named *WikiMir*. The system seeks to retrieve mathematical information based on keywords, and the structure and importance of formulae in a document. Furthermore, the system proposes a novel hybrid indexing and matching model to support exact and fuzzing matching. In this hybrid model, both keyword and structure information of formulae are taken into consideration. In addition, the concept of formula importance within a document is introduced into the model. In order to make the results more reasonable, the system re-ranks the top-k formulae by regular expressions matching of the query formula.

Table 4: NTCIR-12 MathIR Task Summary.

Group ID	Organization		No. Runs by Task: arXiv-main (14)	arXiv-simto (8)	Wiki-main (18)	Wiki-formula (7)
ICST	Peking U. Inst. of Comp. Sci. & Tech.	(CN)	1		1	
FSE	TU Berlin & University of Konstanz	(DE)			1	
MCAT*	National Institute of Informatics	(JP)	4	4	4	3
MIRMU	Masaryk University	(CZ)	4	4	4	
RITUW*	Rochester Inst. Tech. & Univ. Waterloo	(US, CA)	4		4	4
SMSG5	Samsung R&D India-Bangalore	(IN)	1		4	

*Task organizers

Table 5: Participant System Configurations. All systems support query keywords and formulae. Center columns relate to formulae; the rightmost column indicates whether existing search engines are used.

RunID	Formula Encodings Used: LaTeX	Presentation	Content	Tree Structure	Query Variables	Search Engine
$MCAT_{\text{ND-LR-U}}$	No	Yes	Yes	Yes	Yes	Yes
$MCAT_{\text{A-NW-U}}$					Yes	
$MCAT_{\text{A-LR}}$					No	
$MCAT_{\text{A-LR-U}}$					Yes	
FSE_{RUN1}	No	Yes	No	Yes	Yes	Yes
ICST	No	Yes	No	Yes	Yes	Yes
RITUW (all runs)	No	Yes	No	Yes	Yes	Yes[1]
$MIRMU_{\text{CM-R-10}}$	No	No	Yes	Yes	No	Yes
$MIRMU_{\text{PM-R-10}}$		Yes	No			
$MIRMU_{\text{PCM-L-19}}$		Yes	Yes			
$MIRMU_{\text{PM-L-19}}$		Yes	No			
$MIRMU_{\text{PCM-L-18}}$		Yes	Yes			
$MIRMU_{\text{CM-L-19}}$		No	Yes			
$MIRMU_{\text{CM-R-19}}$		No	Yes			
$SMSG5_{\text{TFIDF}}$	No	Yes	No	No	Yes	Yes

*1 Text only

Table 6: Relevance Assessment. Most hits were rated independently by two evaluators using the Sepia interface (see Figure 1), and the final rating was the sum of their scores (*Combined*). A few hits in the arXiv tasks had 3-4 evaluators; here the average individual rating was doubled and rounded down to produce the final combined rating.

Assessment	Individual	Combined
Relevant	2	3-4
Partially Relevant	1	1-2
Non Relevant	0	0

FSE (TU Berlin and Univ. Konstanz [11])

The FSE team used a simple method to create a manual run for the Wiki-main task. The primary author, a physicist and computer scientist looked at the queries and entered the titles of associated Wikipedia pages in the search interface at `en.wikipedia.org`. For some topics, the German Wikipedia version was used first, and inter-language links were used to identify a corresponding English Wikipedia page. In a second step, which consumed the most time, the team identified the corresponding documents in the dump. For some hits, we were unable to find a corresponding document in the Wikipedia corpus.

MCAT (National Institute of Informatics [7])

The MCAT group implemented an indexing scheme for math expressions within an Apache Solr database.

Innovations include three levels of granularity for textual information (math, paragraph, and document levels), a method for extracting dependency relationships between math expressions, score normalization, cold-start weights, and unification. Dependency relationships and unification improved search precision significantly. The cold start weights, however, did not have a good impact on the search performance, perhaps due to the negative weights obtained for several fields in their database. On the other hand, the applied score normalization worked well, allowing the system to utilize many fields for search without concern for database fields improperly dominating the final similarity score.

MIRMU (Masaryk University [10])

The Masaryk University Math Information Retrieval (MIRMU) team used their *MIaS* system [14] to participate in the arXiv-main and Wiki-main tasks. Using the NTCIR-11 Math-2 Task relevance judgements [2], an evaluation platform was developed [8] to rigorously evaluate combinations of new features, and then select the most promising ones for the NTCIR-12 evaluation.

New features were aimed primarily at further canonicalizing MathML input, structural unification of formulae for syntactic-based similarity search, and query expansion to obtain better results for combined text and math queries.

RITUW (RIT and University of Waterloo [3])

The Tangent-3 system uses two indices: 1) a Solr-based index for document text, and 2) a custom inverted index for math expressions. Pairs of symbols along with their spatial relationships in Presentation MathML define tokens for fast lookup in the math expression index (median retrieval time

Table 7: Inter-Rater Agreement for MathIR Tasks. Fleiss' κ is used to measure agreement; *docs* is the number of documents rated by two evaluators; articles rated by three or four evaluators are *skipped*.

	arXiv -main	arXiv -simto	Wiki -main	Wiki -formula
docs (skipped)	4234 (17)	612 (42)	4107	2687
κ	0.5615	0.5380	0.3546	0.2619

of 1.07s for Wiki-formula). Formula candidates returned from the expression index are re-scored, taking structural constraints, wildcard expansion, and symbol unification into account. Text and math indices are queried separately; documents are ranked by a linear combination of math expression and keyword matching (Solr) scores.

Equal weights for keyword match and formulae match scores, and equal weighting for formulae in a query worked best. Constraining unification would improve formula retrieval, along with refined similarity metrics that better exploit the high recall for formulae returned from the index.

SMSG5 (Samsung R&D India-Bangalore [15])

SMSG5 group's main focus was to utilize the co-occurrence of formulae and text to produce more relevant results. This has been done in the past through pattern-based and other approaches. In their approach, they exploit LDA and doc2vec's co-occurrence finding techniques, in addition to pattern and Elastic search-based document ranking. Additionally, a technique is used to merge results from knowledge bases with different scoring mechanisms, using a nested Borda Count-based technique. The resulting re-ranking mechanism is simple and fast.

For the main arXiv task, SMSG5 submitted one run corresponding to Elastic Search (ES) output only. Originally three additional runs were planned: ES + Doc2Vec and ES + Doc2Vec + LDA + Pattern, but due to time constraints (SMSG5 entered late) and an unavailability of appropriate infrastructure, only one run was submitted. For the Wikipedia-main task, four runs were submitted: ES only, ES + Doc2Vec, ES + Doc2Vec + Pattern and ES + Doc2Vec + Pattern + LDA.

4. TASK RESULTS AND DISCUSSION

Relevance Assessments. The distribution of relevance scores for each topic is summarized in Figure 2. Based on the percentage of pooled hits rated as relevant or partially relevant, the Wikipedia tasks appear to have been easier, particularly the formula browsing task. In one case (Topic 9 in the Wiki-main task), the target document for a navigational query was accidentally omitted from the corpus, and so no hits were relevant. Topic 19 in both the arXiv-main and Wiki-main tasks had very narrow information needs, with few relevant hits.

Table 7 shows Fleiss' κ for each task. This statistic is used to measure agreement between assessors. Agreement between evaluators for the arXiv tasks is higher. This may be because of the greater mathematical expertise and shared background by these evaluators. For the Wikipedia task, we observed informally that the undergraduate evaluators frequently rated hits differently than the Master's students, who had often studied more mathematics. It is also interesting that the Wiki-formula task has the lowest agreement value, despite the very high percentage of partially relevant hits in Figure 2. We observed some evaluators were very concerned with formula semantics, while others seemed to consider primarily visual similarity when rating hits.

Performance Metrics. Table 8 shows the results for all runs. Performance metrics are averaged over all the queries. Note that these percentages can be misleading; 60% at top-5 corresponds to three hits, but at top-20 corresponds to 12. Some teams worked under a misunderstanding that rank and not score would be used to order hits for evaluation; for their benefit, in Table 9 we provide unofficial results calculated by substituting inverse rank for score.

For all but the arXiv-simto task, metrics for the pool are provided, obtained by sorting all of the pooled top-20 hits in decreasing order by relevance rating. In most cases, there is a substantial gap between this ideal result taken from the pool and the strongest result for individual runs.

Discussion. The best-performing systems (MCAT and ICST) utilize textual context for formulae, and integrate retrieval of text and formulae. ICST performed much more strongly in the Wiki-main task than the arXiv-main task, perhaps because the full Wikipedia articles contain more text, and/or because the navigation structure of the encyclopedia is used, weighting articles based on the number of links to and from an article, similar to PageRank [9].

In the unofficial rank-based results the RITUW system obtains competitive results for the arXiv-main task (e.g., obtaining the second-highest Precision@5), but then performs more weakly than runs from ICST, MCAT and SMSG5 in the Wiki-main task, as all of these systems integrate text and formula retrieval.

The manual FSE run for the Wiki-main task identified additional hits not located by the automated systems, enriching the pool. FSE provide a detailed record of hits they identified for each query in the Wiki-main task, along with an analysis of shared links in relevant articles [11].

In the arXiv-simto task, the MCAT system's support for query variables and context may have led to stronger results than the MIRMU runs. In the Wiki-formula task, MCAT slightly outperforms RITUW, but with slower retrieval times. MCAT uses Presentation and Content MathML formula encodings along with textual context etc.; RITUW uses only Presentation MathML.

Unification appears to be beneficial for re-ranking, but slows systems down. Unification for candidates with few matching symbols appears to hurt precision.

5. CONCLUSION

The NTCIR-12 MathIR task is the third Math Information Retrieval (MIR) task at an international IR evaluation forum. A new test collection of Wikipedia articles has been created, along with search topics based on mathematically sophisticated users (arXiv-main) as well as topics reflecting the information needs of mathematical non-experts (Wiki-main). Two other tasks were carried out, one experimental task exploring a new query language operation (arXiv-simto) along with a formula browsing task (Wiki-formula). Our arXiv and Wikipedia corpora are available, and topics and assessment ratings will be released later in 2016.

Differences between ideal rankings from the pools and best individual runs are substantial. Some refinement and combination of techniques might be used to bridge this gap.

A standing question about whether appearance-based or semantics-based encodings are more effective for formula retrieval remains. However, the outcome of our task suggests that this may no longer be an interesting question - the best results were obtained using formula appearance *and* semantics, and perhaps this combination is the right way forward.

Acknowledgments

The work reported here here has been partially supported by the Leibniz association under grant SAW-2012-FIZ_KA-2, JSPS KAKENHI Grant Numbers 2430062, and CREST, JST, and the National Science Foundation (USA) under grant no. HCC-1218801. We are grateful to Kazuki Hayakawa and Takeshi Sagara for assisting with the task organization, Deyan Ginev for creating the arXiv corpus, Michal Růžička for generating XHTML files for the Wikipedia corpus, and Anurag Agarwal for assistance with designing the Wikipedia queries. Finally, we thank the students who evaluated the search hit pools at Jacobs University (arXiv) and RIT (Wikipedia).

6. REFERENCES

[1] A. Aizawa, M. Kohlhase, and I. Ounis. NTCIR-10 Math pilot task overview. In N. Kando and K. Kishida, editors, *NTCIR Workshop 10 Meeting*, pages 1–8, Tokyo, Japan, 2013.

[2] A. Aizawa, M. Kohlhase, I. Ounis, and M. Schubotz. NTCIR-11 math-2 task overview. In *NTCIR*. National Institute of Informatics (NII), 2014.

[3] K. Davila, R. Zanibbi, A. Kane, and F. Tompa. Tangent-3 at the NTCIR-12 MathIR task. In *Proc. NTCIR-12*, 2016.

[4] L. Gao, K. Yuan, Y. Wang, Z. Jiang, and Z. Tang. The math retrieval system of ICST for NTCIR-12 MathIR task. In *Proc. NTCIR-12*, 2016.

[5] F. Guidi and C. S. Coen. A survey on retrieval of mathematical knowledge. In M. K. et al., editor, *Proc. CICM*, volume 9150 of *LNAI*, pages 296–315, 2015.

[6] M. Kohlhase. Formats for topics and submissions for the MathIR task at NTCIR-12. Technical report, NTCIR, 2015.

[7] G. Y. Kristianto, G. Topić, and A. Aizawa. MCAT math retrieval system for NTCIR-12 MathIR task. In *Proc. NTCIR-12*, 2016.

[8] M. Líška, P. Sojka, and M. Růžička. Combining text and formula queries in math information retrieval: Evaluation of query results merging strategies. In *Proc. Int'l Work. Novel Web Search Interfaces and Systems*, pages 7–9, New York, 2015.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[10] M. Růžička, P. Sojka, and M. Líška. Math indexer and searcher under the hood: Fine-tuning query expansion and unification strategies. In *Proc. NTCIR-12*, 2016.

[11] M. Schubotz, M. Leich, N. Meuschke, and B. Gipp. Exploring the one-brain barrier: a manual contribution to the NTCIR-12 Math task. In *Proc. NTCIR-12*, 2016.

[12] M. Schubotz, A. Youssef, V. Markl, and H. S. Cohl. Challenges of mathematical information retrieval in the NTCIR-11 Math Wikipedia Task. In *Proc. ACM SIGIR*, pages 951–954, 2015.

[13] sepia: Standard evaluation package for information access systems. `https://code.google.com/p/sepia/`.

[14] P. Sojka and M. Líška. The Art of Mathematics Retrieval. In *Proc. ACM DocEng*, pages 57–60, Mountain View, CA, Sept. 2011.

[15] A. Thanda, A. Agarwal, K. Singla, A. Prakash, and A. Gupta. A document retrieval system for math queries. In *Proc. NTCIR-12*, 2016.

[16] Common evaluation measures. In E. M. Voorhees and L. P. Buckland, editors, *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings*, number SP 500-274 in NIST Special Publication, 2007.

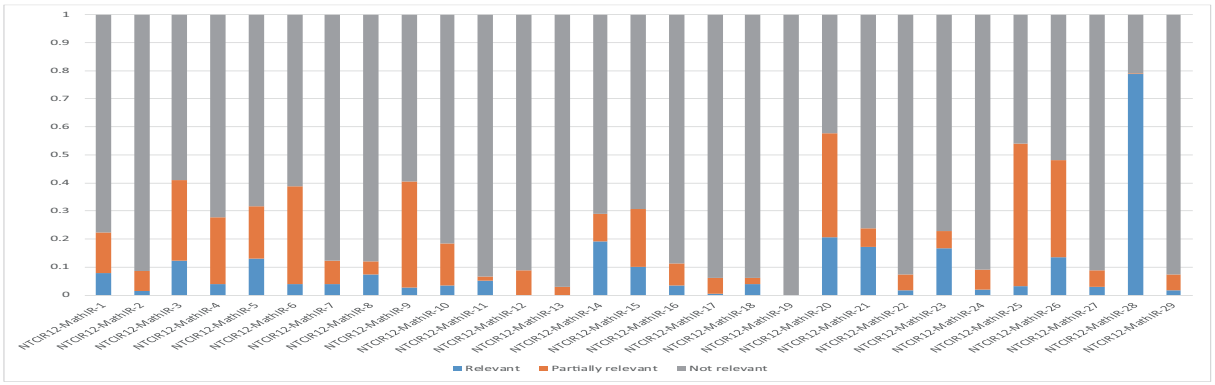
[17] R. Zanibbi and D. Blostein. Recognition and retrieval of mathematical expressions. *Int. J. Doc. Analysis and Recognition*, 15(4):331–357, 2012.

APPENDIX

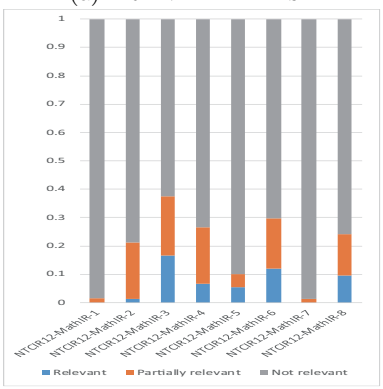
A. TOOLS

The following tools may be useful for this task.

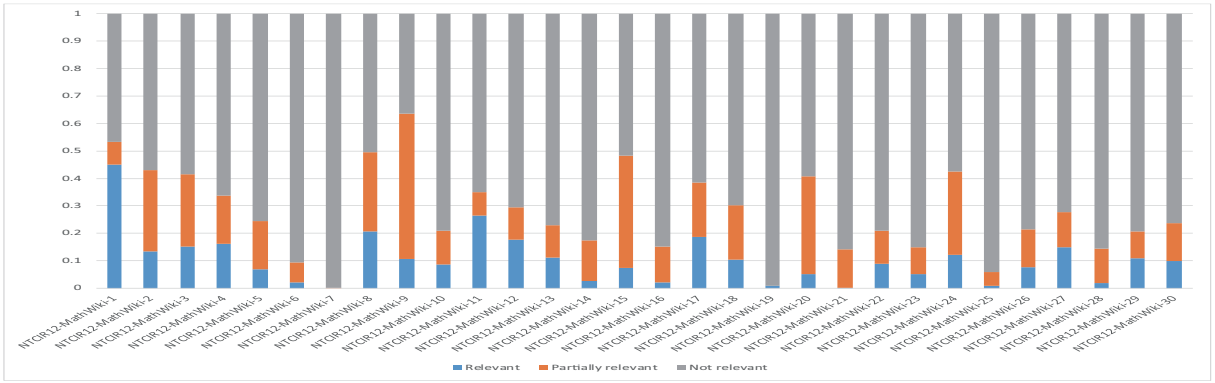
- SEPIA: Standard Evaluation Package for Information Access Systems. Used with MathML extension. (`https://code.google.com/p/sepia/`)
- trec_eval: A program to evaluate TREC results. (`http://trec.nist.gov/trec_eval/`)
- MathJax : Javascript LaTeX/MathML rendering. (`http://www.mathjax.org/`)
- LaTeXML: A LaTeX to MathML converter. (`http://dlmf.nist.gov/LaTeXML/`)
- docs2harvest: Parses html / xhtml documents and generates harvest files with Content Math data only. (`https://github.com/KWARC/mws`)
- mathml−converter: Converts MathML into keywords. (`http://code.google.com/p/mathml-converter/`)



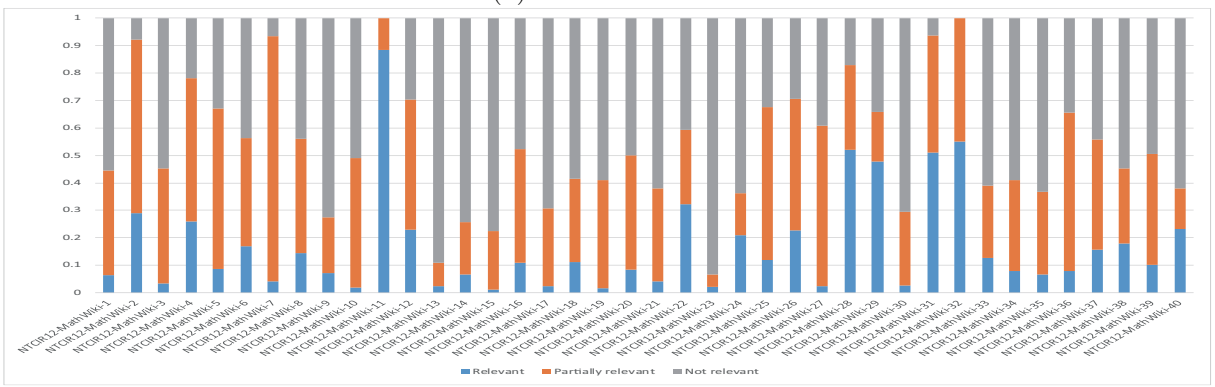
(a) arXiv Main Task



(b) arXiv Simto Task



(c) MathWiki Task



(d) MathWikiFormula Task

Figure 2: Relevance Assessment Statistics. For each query, the percentage of Relevant (blue), Partially Relevant (orange) and Non Relevant (grey) hits in the pool of top-20 hits returned by participant systems are shown. Note that different queries have different pool sizes. Table 6 explains how hit ratings are assigned.

Table 8: Retrieval Performance Summary (Official `treceval`). From left-to-right, Precision@5, 10, 15 and 20 are shown for relevant (rating 3-4), and then relevant and partially relevant hits (rating 1-4). *Pool* shows results for all pooled hits sorted in decreasing order by relevance rating.

Run	Relevant				Partially Relevant			
arXiv Main Task								
$ICST_{\text{main}}$	0.2276	0.1862	0.1632	0.1362	0.5517	0.4966	0.4299	0.4000
$MCAT_{\text{af-lr}}$	0.2552	0.2379	0.2092	0.1828	**0.5586**	0.5379	0.5034	0.4690
$MCAT_{\text{af-lr-u}}$	0.2621	**0.2448**	0.2046	0.1810	**0.5586**	**0.5483**	0.5126	0.4707
$MCAT_{\text{af-nw-u}}$	**0.2828**	0.2379	**0.2184**	**0.1948**	0.5448	0.5345	**0.5149**	**0.4897**
$MCAT_{\text{nd-lr-u}}$	0.2345	0.1966	0.1747	0.1586	0.4828	0.4793	0.4690	0.4552
$MIRMU_{\text{cm-r-10}}$	0.1241	0.1345	0.1218	0.1069	0.3931	0.3655	0.3402	0.3207
$MIRMU_{\text{pm-r-10}}$	0.1172	0.0828	0.0897	0.0810	0.3379	0.2690	0.2943	0.2672
$MIRMU_{\text{pcm-l-19}}$	0.1310	0.1000	0.0782	0.0845	0.3862	0.3483	0.2989	0.2793
$MIRMU_{\text{pm-l-19}}$	0.0690	0.0793	0.0736	0.0793	0.2897	0.2897	0.2644	0.2586
$RITUW_1$	0.2069	0.1517	0.1126	0.0948	0.4966	0.3966	0.3310	0.2879
$RITUW_2$	0.2069	0.1517	0.1126	0.0948	0.4966	0.3966	0.3310	0.2879
$RITUW_3$	0.1379	0.1138	0.1034	0.0914	0.4897	0.4586	0.4207	0.3983
$RITUW_4$	0.1379	0.1138	0.1034	0.0914	0.4897	0.4586	0.4207	0.3983
$SMSG5_{\text{tfidf}}$	0.0690	0.0931	0.0874	0.0810	0.3517	0.3724	0.3586	0.3397
Pool	*0.6966*	*0.5586*	*0.4644*	*0.4086*	*0.9655*	*0.96662*	*0.9172*	*0.8828*
arXiv Simto Task								
$MCAT_{\text{s-af-lr}}$	0.1750	0.1375	0.1417	0.1313	0.4250	0.3875	**0.3833**	**0.3687**
$MCAT_{\text{s-af-lr-u}}$	**0.2750**	0.1625	0.1500	0.1313	0.5000	0.3625	0.3333	0.3063
$MCAT_{\text{s-af-nw-u}}$	**0.2750**	**0.2125**	**0.1667**	**0.1375**	**0.5500**	**0.4250**	0.3667	0.3250
$MCAT_{\text{s-nd-lr-u}}$	0.2000	0.1125	0.0917	0.0687	0.3250	0.2000	0.1667	0.1500
$MIRMU_{\text{pcm-l-18}}$	0.1250	0.0625	0.0500	0.0375	0.2500	0.1750	0.1583	0.1625
$MIRMU_{\text{cm-l-19}}$	0.0250	0.0625	0.0750	0.0625	0.3000	0.2500	0.2667	0.2813
$MIRMU_{\text{cm-r-19}}$	0.0500	0.0750	0.0667	0.0563	0.2750	0.3000	0.2250	0.2063
$MIRMU_{\text{pcm-l-19}}$	0.0500	0.0625	0.0667	0.0563	0.3000	0.2625	0.2417	0.2563
MathWiki Task								
FSE_{run1}	0.1733	0.0867	0.0578	0.0433	0.2333	0.1167	0.0778	0.0583
$ICST_{\text{w-main}}$	**0.4733**	**0.3767**	**0.2978**	**0.2617**	**0.8533**	**0.7900**	**0.7133**	**0.6600**
$MCAT_{\text{af-lr}}$	0.2467	0.2233	0.2044	0.1817	0.5533	0.5133	0.4956	0.4650
$MCAT_{\text{af-lr-u}}$	0.2867	0.2533	0.2244	0.1983	0.6067	0.5700	0.5533	0.5183
$MCAT_{\text{af-nd-lr-u}}$	0.2933	0.2467	0.2267	0.1983	0.6200	0.5867	0.5711	0.5333
$MCAT_{\text{af-nw-u}}$	0.3600	0.3233	0.2689	0.2433	0.7667	0.7167	0.6867	0.6533
$MIRMU_{\text{pm-l-20}}$	0.0600	0.0433	0.0356	0.0333	0.2933	0.2367	0.2222	0.2050
$MIRMU_{\text{pm-r-20}}$	0.0533	0.0400	0.0356	0.0333	0.2933	0.2600	0.2378	0.2167
$MIRMU_{\text{pcm-l-29}}$	0.0600	0.0533	0.0444	0.0400	0.2000	0.2000	0.1867	0.1833
$MIRMU_{\text{pcm-r-29}}$	0.0533	0.0500	0.0444	0.0433	0.2533	0.2167	0.2089	0.1950
$RITUW_{\text{w-b}}$	0.0267	0.0267	0.0267	0.0250	0.1533	0.1567	0.1444	0.1400
$RITUW_{\text{w-1}}$	0.2533	0.2367	0.2089	0.1983	0.4933	0.4833	0.4889	0.4733
$RITUW_{\text{w-2}}$	0.2533	0.2467	0.2156	0.2017	0.4933	0.4900	0.4822	0.4700
$RITUW_{\text{w-3}}$	0.1600	0.1300	0.1244	0.1267	0.3867	0.3633	0.3733	0.3617
$RITUW_{\text{w-4}}$	0.1600	0.1400	0.1311	0.1300	0.3800	0.3667	0.3644	0.3583
$SMSG5_{\text{es}}$	0.3067	0.2567	0.2111	0.1950	0.7533	0.7000	0.6733	0.6467
$SMSG5_{\text{esdv}}$	0.3667	0.2667	0.2444	0.2150	0.7067	0.6633	0.6289	0.6150
$SMSG5_{\text{esdvldapat}}$	0.2867	0.2633	0.2333	0.2183	0.6800	0.6667	0.6578	0.6333
$SMSG5_{\text{esdvpat}}$	0.3667	0.2900	0.2400	0.2233	0.7067	0.6733	0.6511	0.6250
Pool	*0.8400*	*0.6967*	*0.5956*	*0.5133*	*0.9467*	*0.9400*	*0.9289*	*0.9217*
MathWikiFormula Task								
$MCAT_{\text{f-af-lr}}$	0.4100	0.3225	0.2733	0.2325	0.7700	0.7375	0.7083	0.6650
$MCAT_{\text{f-af-lr-u}}$	0.4550	0.3475	0.2950	0.2613	0.7850	0.7475	0.7100	0.6700
$MCAT_{\text{f-af-nw-u}}$	**0.4900**	**0.3900**	**0.3317**	**0.2825**	**0.9100**	**0.8400**	**0.8067**	**0.7687**
$RITUW_{\text{f-1}}$	0.4150	0.3150	0.2650	0.2200	0.8100	0.7450	0.7117	0.6737
$RITUW_{\text{f-2}}$	0.4250	0.3175	0.2567	0.2200	0.8150	0.7550	0.7200	0.6938
$RITUW_{\text{f-3}}$	0.4400	0.3225	0.2700	0.2300	0.8400	0.7650	0.7317	0.7063
$RITUW_{\text{f-4}}$	0.4450	0.2925	0.2517	0.2200	0.8250	0.6825	0.6533	0.6100
Pool	*0.7900*	*0.6400*	*0.5385*	*0.4725*	*1.000*	*1.0000*	*0.9933*	*0.9800*

Table 9: Retrieval Performance Summary (Provided Ranks). From left-to-right, Precision@5, 10, 15 and 20 are shown for relevant (rating 3-4), and then relevant and partially relevant hits (rating 1-4). *Pool* shows results for all pooled hits sorted in decreasing order by relevance rating.

Run	Relevant				Partially Relevant			
arXiv Main Task (*by rank)								
$ICST_{\text{main}}$	0.2207	0.1828	0.1609	0.1379	0.5379	0.4931	0.4437	0.4172
$MCAT_{\text{af-lr}}$	0.2552	0.2379	0.2092	0.1845	0.5586	0.5379	0.5080	0.4810
$MCAT_{\text{af-lr-u}}$	0.2621	**0.2448**	0.2092	0.1845	0.5586	0.5483	0.5218	0.4931
$MCAT_{\text{af-nw-u}}$	**0.2897**	**0.2448**	**0.2276**	**0.2000**	**0.5793**	**0.5552**	**0.5402**	**0.5121**
$MCAT_{\text{nd-lr-u}}$	0.2345	0.2000	0.1793	0.1621	0.4828	0.4793	0.4828	0.4759
$MIRMU_{\text{cm-r-10}}$	0.1241	0.1345	0.1218	0.1069	0.3931	0.3690	0.3425	0.3224
$MIRMU_{\text{pm-r-10}}$	0.1172	0.0828	0.0897	0.0810	0.3379	0.2690	0.2943	0.2672
$MIRMU_{\text{pcm-l-10}}$	0.1310	0.1000	0.0782	0.0845	0.3862	0.3483	0.2989	0.2793
$MIRMU_{\text{pm-l-10}}$	0.0690	0.0793	0.0736	0.0793	0.2897	0.2897	0.2644	0.2586
$RITUW_1$	0.2552	0.2000	0.1586	0.1345	0.5517	0.4517	0.3908	0.3483
$RITUW_2$	0.2621	0.2000	0.1632	0.1362	0.5448	0.4552	0.3908	0.3517
$RITUW_3$	0.1862	0.1552	0.1425	0.1259	0.5448	0.4931	0.4575	0.4414
$RITUW_4$	0.1862	0.1586	0.1425	0.1276	0.5310	0.5034	0.4644	0.4448
$SMSG5_{\text{tfidf}}$	0.0690	0.0931	0.0874	0.0810	0.3517	0.3724	0.3586	0.3397
Pool	*0.6966*	*0.5586*	*0.4644*	*0.4086*	*0.9655*	*0.96662*	*0.9172*	*0.8828*
arXiv Simto Task (*by rank)								
$MCAT_{\text{s-af-lr}}$	0.1750	0.1375	0.1417	0.1313	0.4250	0.3875	**0.3833**	**0.3687**
$MCAT_{\text{s-af-lr-u}}$	**0.2750**	0.1625	0.1500	0.1313	0.5000	0.3625	0.3333	0.3063
$MCAT_{\text{s-af-nw-u}}$	**0.2750**	**0.2125**	**0.1667**	**0.1438**	**0.5500**	**0.4250**	0.3667	0.3312
$MCAT_{\text{s-nd-lr-u}}$	0.2000	0.1125	0.0917	0.0750	0.3250	0.2000	0.1667	0.1562
$MIRMU_{\text{pcm-l-18}}$	0.1250	0.0625	0.0500	0.0375	0.2500	0.1750	0.1583	0.1625
$MIRMU_{\text{cm-l-19}}$	0.0250	0.0625	0.0750	0.0625	0.3000	0.2500	0.2667	0.2813
$MIRMU_{\text{cm-r-19}}$	0.0500	0.0750	0.0667	0.0563	0.2750	0.3000	0.2250	0.2063
$MIRMU_{\text{pcm-l-19}}$	0.0500	0.0625	0.0667	0.0563	0.3000	0.2625	0.2417	0.2563
MathWiki Task (*by rank)								
FSE_{run1}	0.1733	0.0867	0.0578	0.0433	0.2333	0.1167	0.0778	0.0583
$ICST_{\text{w-main}}$	**0.4733**	**0.3767**	**0.2978**	**0.2617**	**0.8533**	**0.7900**	**0.7133**	**0.6600**
$MCAT_{\text{af-lr}}$	0.2467	0.2233	0.2089	0.1867	0.5533	0.5167	0.5067	0.4750
$MCAT_{\text{af-lr-u}}$	0.2867	0.2533	0.2222	0.1933	0.6067	0.5733	0.5556	0.5167
$MCAT_{\text{af-nd-lr-u}}$	0.2867	0.2500	0.2289	0.1950	0.6133	0.5900	0.5733	0.5367
$MCAT_{\text{af-nw-u}}$	0.3600	0.3233	0.2689	0.2433	0.7667	0.7167	0.6867	0.6533
$MIRMU_{\text{pm-l-20}}$	0.0600	0.0433	0.0356	0.0333	0.3000	0.2400	0.2289	0.2117
$MIRMU_{\text{pm-r-20}}$	0.0533	0.0400	0.0356	0.0333	0.2933	0.2600	0.2378	0.2167
$MIRMU_{\text{pcm-l-29}}$	0.0600	0.0533	0.0444	0.0400	0.2000	0.2000	0.1867	0.1833
$MIRMU_{\text{pcm-r-29}}$	0.0533	0.0500	0.0444	0.0433	0.2533	0.2167	0.2089	0.1950
$RITUW_{\text{w-b}}$	0.0600	0.0533	0.0511	0.0500	0.1933	0.1900	0.1733	0.1717
$RITUW_{\text{w-1}}$	0.2467	0.2333	0.2156	0.2050	0.4933	0.4900	0.5000	0.4850
$RITUW_{\text{w-2}}$	0.2533	0.2500	0.2200	0.2050	0.4933	0.4933	0.4867	0.4767
$RITUW_{\text{w-3}}$	0.1600	0.1267	0.1222	0.1250	0.3867	0.3667	0.3689	0.3567
$RITUW_{\text{w-4}}$	0.1533	0.1400	0.1289	0.1250	0.3800	0.3667	0.3600	0.3550
$SMSG5_{\text{es}}$	0.3067	0.2567	0.2111	0.1950	0.7533	0.7000	0.6733	0.6467
$SMSG5_{\text{esdv}}$	0.3667	0.2667	0.2444	0.2150	0.7067	0.6633	0.6289	0.6150
$SMSG5_{\text{esdvldapat}}$	0.2867	0.2633	0.2333	0.2183	0.6800	0.6667	0.6578	0.6333
$SMSG5_{\text{esdvpat}}$	0.3667	0.2900	0.2400	0.2233	0.7067	0.6733	0.6511	0.6250
Pool	*0.8400*	*0.6967*	*0.5956*	*0.5133*	*0.9467*	*0.9400*	*0.9289*	*0.9217*
MathWikiFormula Task (*by rank)								
$MCAT_{\text{f-af-lr}}$	0.4250	0.3350	0.2850	0.2450	0.7850	0.7550	0.7267	0.6875
$MCAT_{\text{f-af-lr-u}}$	0.4750	0.3675	0.3117	0.2775	0.8050	0.7725	0.7333	0.6963
$MCAT_{\text{f-af-nw-u}}$	**0.5150**	**0.4050**	**0.3450**	**0.3000**	**0.9300**	**0.8650**	**0.8300**	**0.8012**
$RITUW_{\text{f-1}}$	0.4300	0.3400	0.2933	0.2450	0.8400	0.7800	0.7533	0.7225
$RITUW_{\text{f-2}}$	0.4450	0.3675	0.3100	0.2687	0.8550	0.8125	0.7833	0.7638
$RITUW_{\text{f-3}}$	0.4900	0.3750	0.3283	0.2812	0.8750	0.8175	0.7833	0.7563
$RITUW_{\text{f-4}}$	0.4900	0.3750	0.3217	0.2937	0.9000	0.8250	0.8033	0.7762
Pool	*0.7900*	*0.6400*	*0.5385*	*0.4725*	*1.0000*	*1.0000*	*0.9933*	*0.9800*