Preprint from https://www.gipp.com/pub/

M. Schubotz, N. Meuschke, M. Leich, and B. Gipp, "Exploring the One-brain Barrier: a Manual Contribution to the NTCIR-12 Math Task," in Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-12), 2016.

Exploring the One-brain Barrier: a Manual Contribution to the NTCIR-12 MathIR Task

Moritz Schubotz
Database Systems and Information Management Grp.
Technische Universität Berlin
Berlin, Germany
schubotz@tu-berlin.de

Norman Meuschke
Department of Computer and Information Science
University of Konstanz
Konstanz, Germany
norman.meuschke@uni.kn

Marcus Leich
Database Systems and Information Management Grp.
Technische Universität Berlin
Berlin, Germany
marcus.leich@tu-berlin.de

Bela Gipp
Department of Computer and Information Science
University of Konstanz
Konstanz, Germany
bela.gipp@uni.kn

ABSTRACT

This paper compares the search capabilities of a single human brain supported by the text search built into Wikipedia with state-of-the-art math search systems. To achieve this, we compare results of manual Wikipedia searches with the aggregated and assessed results of all systems participating in the NTCIR-12 MathIR Wikipedia Task. For 26 of the 30 topics, the average relevance score of our manually retrieved results exceeded the average relevance score of other participants by more than one standard deviation. However, math search engines at large achieved better recall and retrieved highly relevant results that our 'single-brain system' missed for 12 topics. By categorizing the topics of NTCIR-12 into six types of queries, we observe a particular strength of math search engines to answer queries of the types 'definition look-up' and 'application look-up'. However, we see the low precision of current math search engines as the main challenge that prevents their wide-spread adoption in STEM research. By combining our results with highly relevant results of all other participants, we compile a new gold standard dataset and a dataset of duplicate content items. We discuss how the two datasets can be used to improve the query formulation and content augmentation capabilities of match search engines in the future.

Team and Task Information

Team Name: Formula Search Engine (FSE)
Task: optional MathIR Wikipedia Task (English)

Keywords

Math Search, MathML, Manual Contribution

1. INTRODUCTION

The flexiformalist manifesto [7] describes the vision to use machine support to break the 'one-brain barrier', i.e. to combine the strengths of humans and machines to advance mathematics research. The one-brain barrier is a metaphor describing that to solve a mathematical problem all relevant knowledge must be co-located in a single-brain. The past and current NTCIR MathIR tasks [1, 2, 3] and the MIR Happening at CICM'12 defined topics supposed to model mathematical information needs of a human user. These events have welcomed submissions generated using information retrieval systems as well as submissions that domain experts compiled manually. However, there have not been any manual submissions to these events so far.

Having submitted machine-generated results in the past [9, 11, 13], we submitted a manual run for the Wikipedia task of NTCIR-12. Motivated by a user study that analyzed the requirements on math search interfaces [6], we use the manual run to derive insights for refining the query formulation of our math search engine mathosphere and for additional future research on math information retrieval (MIR). Metaphorically speaking, we put human brains in the place of the query interpreter to investigate how well the query language used to formulate the queries of the NTCIR-12 task can convey the respective information needs. We want to see, how the human assessors judge the relevance of results that another human retrieved after interpreting the query compared to the results a machine retrieved.

2. METHODS

2.1 Task overview

The NTCIR-12 optional MathIR Wikipedia Task (English) is the continuation of the NTCIR-11 MathIR Wikipedia task [12]. The participants received 30 search topics, which are ordered lists, whose elements can either be keywords or math elements. For example topic 1 of NTCIR-12 consists of four elements:

1. (Text, `what`)
2. (Text, `symbol`)
3. (Text, `is`)
4. (Math, ζ).

To improve readability, we use a single space to separate the elements of topic descriptions hereafter, i.e. topic 1 reads: 'what symbol is ζ'. The math elements in topic descriptions can include so called `qvar` elements, which are back-referencing placeholders for mathematical sub-expressions.

For example, query nine includes a math expression with one `qvar` element: $*1*_{n+1} = r*1*_n(1-*1*_n)\ *1*$. The `qvar` element can represent any identifier such as x, or a mathematical expression such as $\sqrt{x^2}$, as long as all the occurrences of $*1*$ are replaced with the same sub-expression. Table 4 lists all 30 topics of NTCIR-12. The participants received a subset of the English Wikipedia consisting of: (1) all pages that included a special tag used in MediaWiki to mark up mathematical formulae, hereafter referred as `<math/>`-tag; and (2) a random sample of Wikipedia pages without `<math/>`-tags.

The task was to submit ordered lists of hits for each topic. The results in all submissions (manual and automated) were pooled. We expect that the organizers used a 'page-centric approach' for result pooling, i.e. that hits with the same page title, but pointers to different positions in the page were combined. Due to the pooling process, tracing which engine and how many different engines retrieved a particular result is no longer possible. This aggregation is detrimental to answering our research question, since we cannot evaluate if the results returned by our 'single-brain system' were also found by the math search engines at large.

Two assessors independently evaluated the relevance of each result in the pool using a tripartite scale ranging from 'not relevant = 0' over 'partially relevant = 1' to 'relevant = 2'. As in past NTCIR MathIR tasks, the assessors at NTCIR-12 received relevance criteria as a guideline for their assessments. These relevance criteria have not been made available to the participants prior to the manuscript submission deadline. The anonymized results of the assessment process were distributed to the participants. The task organizers aggregated the assessment scores by summing the two scores given to each result. For future NTCIR MathIR tasks, we propose to provide the individual scores given by each assessor or to state the assessor agreement. The aggregated relevance scores do not allow to deduce assessor agreement with certainty, because of the ambiguous score combinations $2, 0$ and $1, 1$. Details on the topics, the data, the result formats, the dataset, the pooling process, and standard performance measures such as Mean Average Precision are available in [3].

2.2 Pre submission

To generate our results, we followed a simple approach (see Figure 1). We entered the titles of associated Wikipedia pages into the search interface at `en.wikipedia.org`. For some topics, we used the German Wikipedia instead and followed inter-language links to retrieve the corresponding page in the English Wikipedia. Note that our 'single-brain system' was trained in physics and computer science, which might have biased the results. In a second step, we identified the corresponding document in the test collection for the NTCIR-12 task, which was not possible in four cases.

2.3 Post submission

After receiving the relevance feedback of the human assessors, we analyzed all pages that the assessors judged as highly relevant, but our search did not retrieve. We define results that received a score of 4 as highly relevant, i.e. both assessors classified the result as relevant. We used highly relevant results to refine our result list, with the goal of using it as a gold standard for our math search engine mathosphere in the future. Additionally, we generated a list of duplicate results, which we plan to use as training data to improve content augmentation for mathosphere.

Figure 2 shows the distribution of relevance scores for each topic, i.e. how many result pages were assessed per topic and which scores were given. The number of assessed pages ranges from 100 for topic 11 to 178 for topic 28. More interestingly, the number of pages judged as highly relevant varied significantly from 0 for topics 7, 14, 16, and 21 to 49 for topic 1. It appears that some topics were more challenging for the participating search engines than others.

To create the gold standard, we used the following procedure: (1) we added to the result list results that other participants retrieved, but we missed; (2) we re-ranked our result list given the new suggestions of other participants; (3) we removed results that we no longer consider relevant, because they do not add new information to items we ranked higher in our final result list; (4) we excluded topics, for which we considered no result relevant to the query.

During this process, we also tracked duplicates, i.e. content that covers the same information for a topic. While we have not yet fully formalized our notion of duplicate content, we differentiate between two types of duplicates i) parent-child duplicates and ii) sister duplicates. We define 'parent-child duplicates' as a set of duplicate content elements with one distinct element (parent) to which all other elements (children) link. On the contrary, we define 'sister duplicates' as duplicate content elements that do not exhibit a distinctive link pattern between each other.

3. RESULTS

This Section presents the results of our study with regard to (1) our performance in comparison to other participants, (2) the creation of a new gold standard, and (3) the compilation of a dataset of duplicate content. All results we report have been derived from analyzing the official evaluation results distributed to the participants.

Table 1: Assessment matrix for our results. Rows represent the rank we assigned to a result. Columns represent the relevance score a result received from the two assessors.

		relevance score					
		0	1	2	3	4	Σ
rank	1	1	1	4	3	19	28
	2	2	1	1	1	2	7
	3	0	0	0	1	0	1
	4	0	0	1	0	0	1
	5	0	0	1	0	0	1
	Σ	3	2	7	5	21	38

3.1 Performance

For the 30 topics, we retrieved 42 pages that we deemed relevant from `en.wikipedia.org` (see Table 4). Four of our hits (the top hit for topic 7 and the lowest-ranked hits for topic 2, 3, and 13) were not part of the NTCIR-12 corpus. Table 1 shows the relevance assessments of the 38 pages that were part of the corpus. Twenty-one of our results were judged as relevant by both assessors, additional five results were judged as relevant by one and as partially relevant by the other assessor. Of the 28 pages that we considered as

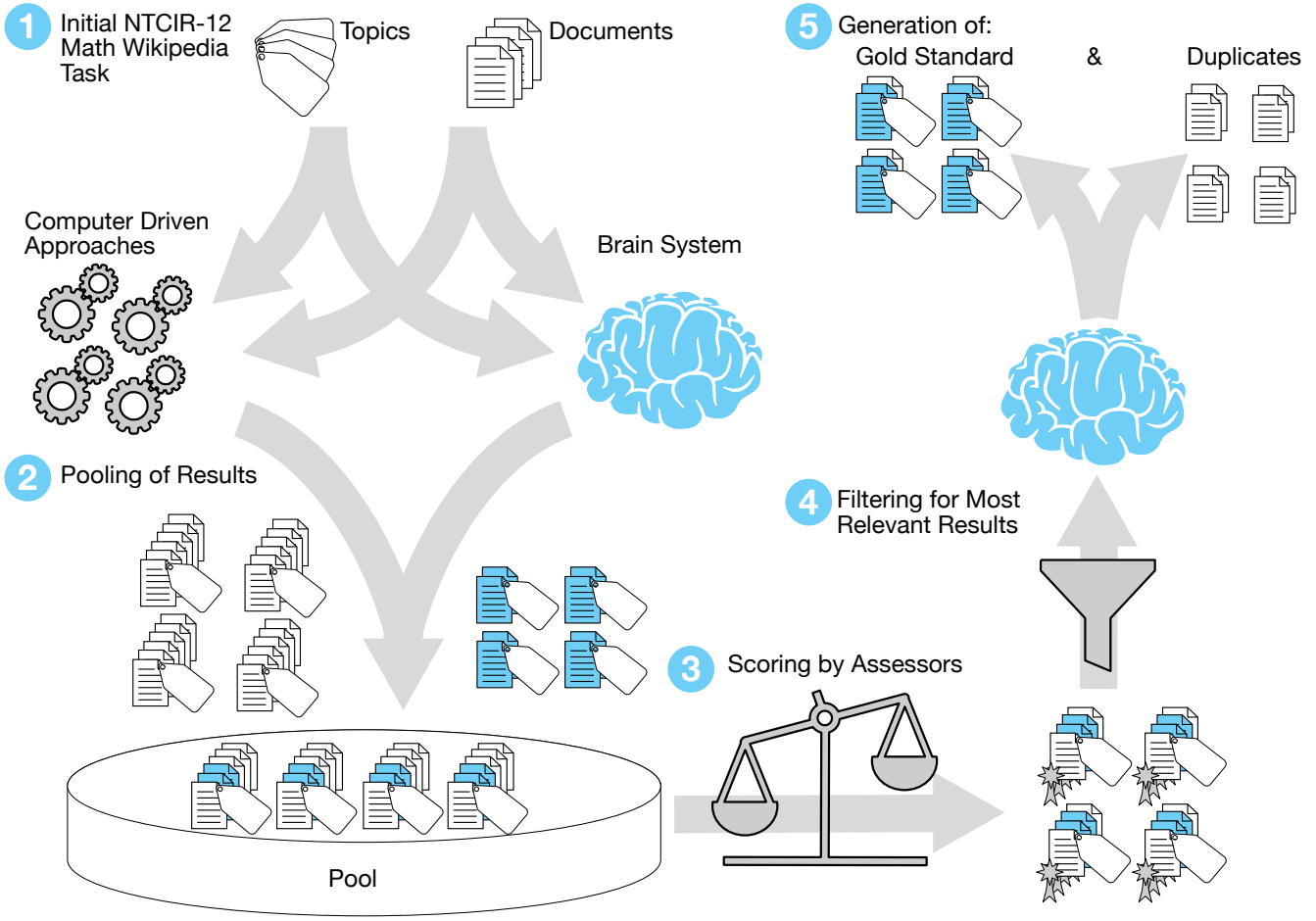

Figure 1: Overview of our experimental setup.

top hits, 19 pages were judged as relevant by both assessors. Only three of our results were judged as irrelevant by both assessors. We could not deduce the reasons that caused the assessors to judge one of our top-ranked results as irrelevant and eight top-ranked results as partially relevant. One explanation could be that the assessors received relevance criteria favoring a perspective that was not apparent from the query. Table 4 explains for each topic whether and why we agreed with the assessors' judgment.

Figure 3 shows for each topic the average relevance score our results received and the average relevance score of all participants. Except for three topics (7, 16 and 21) our scores are clearly above the average for all systems. For 26 of the 30 topics, our average relevance score for the topic exceeded the average score of other participants by more than one standard deviation. The NTCIR-12 organizers report the precision of participating systems at rank 5, 10, 15 and 20. Since we submitted only one result for 24 of the 30 topics and less than five results for 29 of the 30 topics, we achieved a nominally low precision compared to the other participants. We are confident that this low precision is mainly caused by the small number of results we submitted and to a much lesser extent by the number of false positives. We assume that our precision at rank 1 is more competitive to the results of other systems and better represents our true performance. However, the distributed evaluation results lack the necessary detail to quantitatively substantiate this assumption. Interesting additional analyses would be to compare for each topic our top-k results to the top-k results of the best performing system and to the highest rated results across all systems. However, since the official evaluation results exclusively stated aggregated performance measures for other participants, such a topic-specific comparison was unfeasible. The task overview paper [3] will contain a more detailed comparison of our submission to other submissions.

3.2 Gold standard

Since we consider the set of topics as representative for typical queries an average Wikipedia user might have, we generated a gold standard dataset from the topics and results, to train our search engine mathosphere. We exclusively included highly relevant results in our gold standard. Therefore, we excluded two topics (7 and 16) (see Table 5), for which no participant retrieved relevant results. However, we decided to keep topic 21, although the assessors judged our results as irrelevant (see Table 4, Column 21), because we consider our result a good hit for the use cases and information needs our search engine addresses.

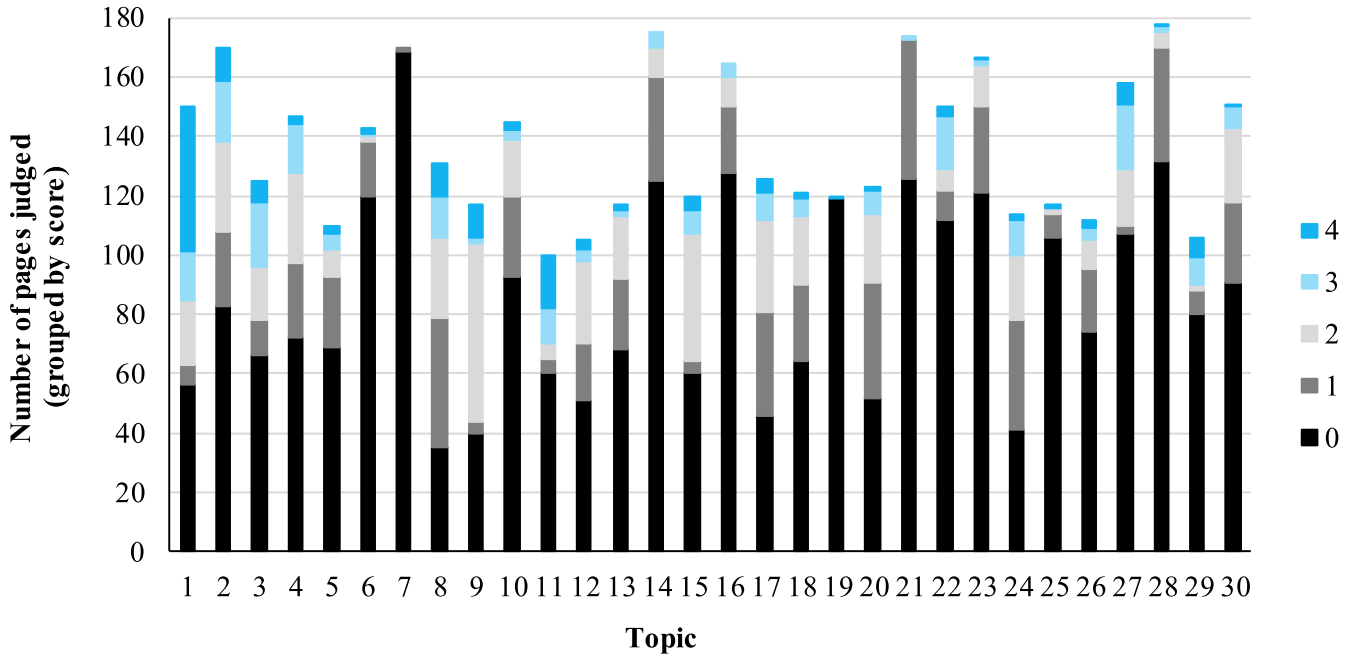

Figure 2: Overview of the 8,214 assessments by topic (4,107 hits, each rated by two reviewers).

To compile the gold standard, we reviewed the 138 results judged as relevant by both assessors and obtained the following results: For 12 of the 28 topics, we found new relevant results (30 in total) that we would not have found without the help of the math search engines participating in NTCIR-12. The search engines may have returned more results that would be beneficial to us, but which we did not review, because they received a score of less than two from either of the assessors. Time constraints caused us to set this strict exclusion criterion.

Finding 30 new results in a set of 138 results pre-classified as relevant might seem like a low yield. However, one has to keep in mind that our goal differs from that of the assessors. While the assessors judged whether the search results are relevant to a specified information need unknown to us, our goal is to decide whether the results are relevant to the query.

For the topics 1 and 27, we discovered the largest numbers of additional relevant results (6 and 7 respectively). These topics reflect two typical types of information needs, i.e. finding definitions of an identifier (topic 1) and finding instantiations of a formula (topic 27). We discussed these types of information needs in [10].

3.3 Duplicates

Table 2: Sister duplicates.

Sister A	Sister B
Logical equivalence	Distributive property
R:K selection theory	Population dynamics
Tautology (rule of inference)	Exportation (logic)

During the evaluation, we identified 60 instances of duplicate results. Most duplicates (57) were parent-child duplicates. For example in the context of topic 8, the Wikipedia article on Wavelength (child) uses the formula $\nu = c/n(0)$ and links to the refractive index (parent) in close proximity to the formula. The other children articles listed in Table 3 exhibit a similar pattern. On the contrary, the Wikipedia article on refractive index, uses the formula $n = c/\nu$ in the abstract and elaborates on this relation throughout the page.

For the sister duplicate relation, we could not identify one article as the main, i.e. parent article. For example in the context of topic 2, the pages "Tautology (rule of inference)" and "Exportation (logic)" contain the exactly same sentence to describe $\Leftrightarrow$: "Where '$\Leftrightarrow$' is a metalogical symbol representing 'can be replaced in a logical proof with'."

4. DISCUSSION

As part of the NTCIR-12 MathIR Wikipedia task, we submitted a set of 38 manually retrieved results. For 17 of the 30 topics, our results achieved a perfect relevance score, for 26 topics our average relevance for the topic exceeded the average relevance score of other participants by more than one standard deviation (see Figure 3).

This outcome demonstrates the strength of our 'single-brain system' and the weaknesses of current math search engines. A human can easily distinguish different query types by analyzing the keywords given in the topic, e.g., retrieving the definition of an identifier unknown to the user opposed to retrieving a complete proof. State-of-the-art math search engines, such as our engine mathosphere, do not yet adapt their search algorithms depending on the keywords given in the topic. Therefore, we see the development of focused math information retrieval algorithms as a promising task for future research. From the current list of topics, we derive the following preliminary set of query types, which correspond to focused MIR tasks:

- Definition look-up (topics 1-3, 5, 6, 19, 24)
- Explanation look-up (topics 20-23)

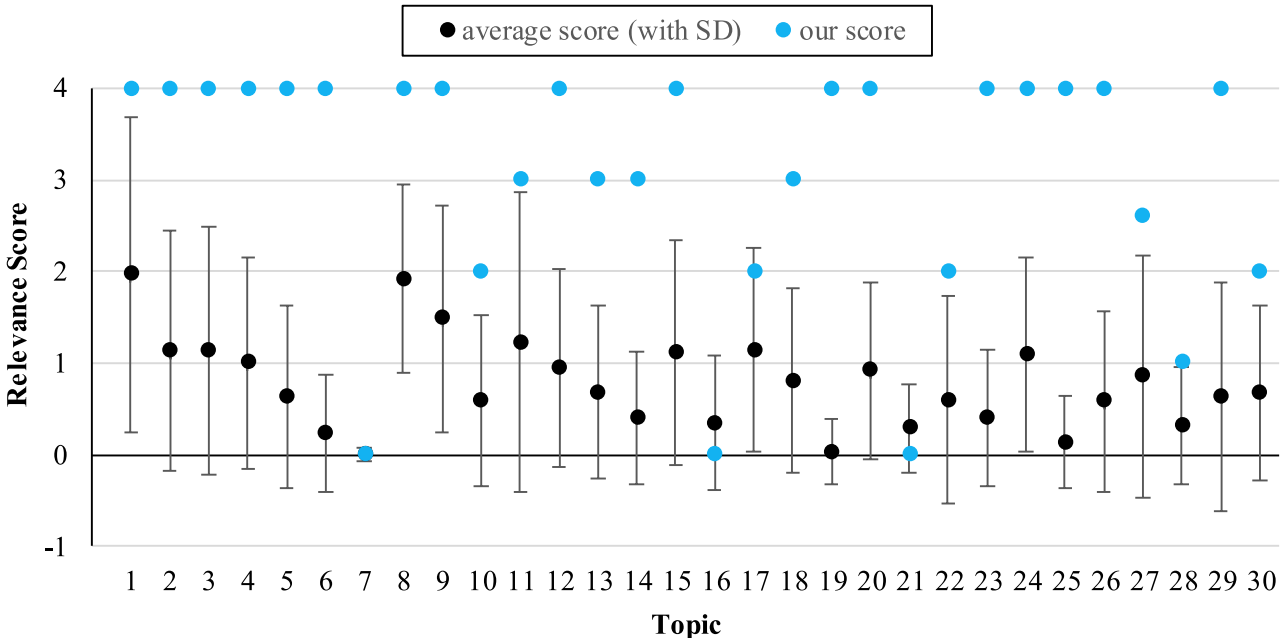

Figure 3: Comparison of our results to the average of the other systems.

- Proof look-up (topics 25, 26)
- Application look-up (topics 27-30)
- Computation assistance (topics 14-18)
- General formula search (topics 4, 7-13)

One approach to address focused MIR tasks is to associate mathematical formulae with corresponding metadata extracted from the mathematical documents and other sources. The NIST Digital Library for Mathematical Functions (DLMF) Project [4, 5] follows this approach. The DLMF offers formulae and their metadata isolated from the formulae's context. Therefore, formulae can serve as standalone retrieval units, since all information necessary to interpret the formulae is given as part of a so called formulae homepages. Extracting high quality metadata is the hardest challenge for this approach. Although research on formulae metadata extraction exists [14, 8, 10], the authors are not aware of search engines that associate metadata with formulae to improve the similarity assessment of formulae. The weakness of our 'single-brain system' is the lower recall compared to math search engines. We submitted only one result for 24 of the 30 topics and less than five results for 29 of the 30 topics. The task organizers report precision at ranks 5, 10, 15 and 20. Given the small number of results we submitted, our precision at rank 5 and above is low. We would have liked to compare for each topic our top-k results to the top-k results of the best performing system and to the highest rated results across all systems. However, these comparisons were unfeasible, since the official evaluation results exclusively stated aggregated performance measures for other participants.

We reviewed all results that both NTCIR assessors rated as relevant, to identify the weaknesses of our 'single-brain system', i.e. which relevant results we missed and why we missed them. Performing this analysis showed that math search engines participating in NTCIR-12 retrieved a number of relevant results, we were unable to find without the support of these engines. This indicates, that the capabilities of today's math search engines to identify relevant formulae exceed the capabilities of humans. However, these super-human capabilities mostly derive from higher recall, while the precision of current math search engines is still low - in our view too low to warrant wide-spread use of these systems by a broad audience.

As we present in Section 3.2, we observed a particular strength of math search engines in answering the queries of topics 1 (definition look-up) and 27 (application look-up). Also for other topics with queries of the types definition look-up (topics 2-5, 24) and application look-up (topics 28-30) math search engines retrieved highly relevant information that that our 'single-brain system' missed.

To train our mathosphere engine to reach a level at which the system could become a widely-used formula search engine for Wikipedia, we compiled a new gold standard dataset (see Table 5). During this process, we additionally created a dataset of duplicate content items (see Tables 2 and 3). We envision to use the duplicate content dataset to enhance the content augmentation phase of the MIR process, i.e. during metadata extraction and index creation, rather than as training data for content querying.

5. CONCLUSION AND OUTLOOK

In conclusion, we regard the NTCIR-12 MathIR Wikipedia task as a valuable preliminary step to develop a formula search engine for Wikipedia. We observed strengths of current math search engines for looking up definitions and applications of formulae. The weakness of current math search engines is their low precision. Improving math search engines to the point where they will become a similarly central fixture for STEM research as keyword-based search engines like Google have become for general purpose queries requires substantial further research and development efforts.

We propose the following steps to reach that goal:

1. Using the gold standard dataset we derived in this paper to train and thereby improve the effectiveness of math search engine prototypes for Wikipedia. The dataset leverages the strengths of the human brain and balances its weaknesses with results of the participating math search engines.

2. Optimizing the efficiency of math search engines as outlined in [11].

3. Generalizing the scope of the improved math search engines from Wikipedia to other collections and more advanced retrieval operations.

Acknowledgments. We thank Volker Markl, Akiko Aizawa and Andrea & Michael Kohlhase for fruitful discussions.

Table 3: Parent-child duplicates.

Parent	**Children**
Binomial coefficient	Binomial theorem, Combination, Lottery mathematics, Pascal's triangle
Damping ratio	RLC circuit
Determinant	Complex number
Direct sum of modules	Exterior algebra, Linear complex structure, Unbounded operator
Faraday's law	Electromagnetic field, Maxwell's equations
Hypergeometric function	Barnes integral, Lauricella hypergeometric series
If and only if	Truth table
Legendre symbol	Jacobi symbol, Quadratic residuosity problem
Logistic function	Feedforward neural network, Sigmoid function, Maximum sustainable yield, Population model, Theoretical ecology
Logistic map	Attractor, Chaos computing, Coupled map lattice, Discrete time and continuous time, File:Cml2e.gif, Parameter space
Mild-slope equation	Sea state
Newton's laws of motion	Braking distance, Cauchy momentum equation, Dynamical simulation, Inertia, Mass in special relativity, Mechanics, Moment of inertia, Newton (unit), Perturbation theory, Pulse wave velocity, Rotation around a fixed axis
Propellant mass fraction	Single-stage-to-orbit
Pythagorean theorem	Pythagoras, Crossed ladders problem, Isosceles triangle, Law of cosines, Slant height, Special right triangles, Triangle
Refractive index	Aether drag hypothesis, Cherenkov radiation, Coherence length, Fizeau experiment, Multiangle light scattering, Optics, Total internal reflection, Wavelength

6. REFERENCES

[1] Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. NTCIR-10 Math Pilot Task Overview. In Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, pages 654–661, Tokyo, Japan, 2013.

[2] Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Moritz Schubotz. NTCIR-11 math-2 task overview. In NTCIR. National Institute of Informatics (NII), 2014.

[3] Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Richard Zanibbi. NTCIR-12 math-3 task overview. In NTCIR. National Institute of Informatics (NII), 2016.

[4] Howard S. Cohl, Marjorie A. McClain, Bonita V. Saunders, Moritz Schubotz, and Janelle C. Williams. Digital repository of mathematical formulae. In Stephen M. Watt, James H. Davenport, Alan P. Sexton, Petr Sojka, and Josef Urban, editors, Intelligent Computer Mathematics - International Conference, CICM 2014, Coimbra, Portugal, July 7-11, 2014. Proceedings, volume 8543 of LNCS, pages 419–422. Springer, 2014.

[5] Howard S. Cohl, Moritz Schubotz, Marjorie A. McClain, Bonita V. Saunders, Cherry Y. Zou, Azeem S. Mohammed, and Alex A. Danoff. Growing the digital repository of mathematical formulae with generic sources. In Manfred Kerber, Jacques Carette, Cezary Kaliszyk, Florian Rabe, and Volker Sorge, editors, Intelligent Computer Mathematics - International Conference, CICM 2015, Washington, DC, USA, July 13-17, 2015, Proceedings, volume 9150 of LNCS, pages 280–287. Springer, 2015.

[6] Andrea Kohlhase. Search interfaces for mathematicians. In Intelligent Computer Mathematics, pages 153–168. Springer, 2014.

[7] Michael Kohlhase. The flexiformalist manifesto. In Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2012 14th International Symposium on, pages 30–35. IEEE, 2012.

[8] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. Extracting textual descriptions of mathematical expressions in scientific papers. D-Lib Magazine, 20(11/12), 2014.

[9] Moritz Schubotz. Making math searchable in wikipedia. CoRR, abs/1304.5475, 2013. DOI: 10.14279/depositonce-5034.

[10] Moritz Schubotz, Alexey Grigoriev, Marcus Leich, Howard S Cohl, Norman Meuschke, Bela Gipp, Abdou S Youssef, and Volker Markl. Semantification of identifiers in mathematics for better math information retrieval. In Proceedings of the 39th international ACM SIGIR conference on Research & development in information retrieval, 2016.

[11] Moritz Schubotz, Marcus Leich, and Volker Markl. Querying large collections of mathematical publications: NTCIR10 math task. In Noriko Kando and Tsuneaki Kato, editors, Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-10, National Center of Sciences, Tokyo, Japan, June 18-21, 2013. National Institute of Informatics (NII), 2013.

[12] Moritz Schubotz, Abdou Youssef, Volker Markl, and Howard S. Cohl. Challenges of mathematical information retrievalin the ntcir-11 math wikipedia task. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, pages 951–954, New York, NY, USA, 2015. ACM.

[13] Moritz Schubotz, Abdou Youssef, Volker Markl, Howard S. Cohl, and Jimmy J. Li. Evaluation of similarity-measure factors for formulae based on the NTCIR-11 math task. In Noriko Kando, Hideo Joho, and Kazuaki Kishida, editors, Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, December 9-12, 2014. National Institute of Informatics (NII), 2014.

[14] Abdou Youssef. Mathematical Knowledge Management: 5th International Conference, MKM 2006, Wokingham, UK, August 11-12, 2006. Proceedings, pages 2–16. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

Listing 1: Use the following BibTeX code to cite this article

```
@InProceedings{Schubotz16a,
  Title     = {Exploring the One-brain Barrier:
               a Manual Contribution to the NTCIR-12 Math Task},
  Author    = {Schubotz, Moritz and Meuschke, Norman and Leich, Marcus and
               Gipp, Bela},
  Booktitle = {Proceedings of the 12th NTCIR Conference on Evaluation of
               Information Access Technologies (NTCIR-12)},
  Year      = {2016}
}
```

Table 4: Submitted results and aggregated relevance score (0-4) in parentheses. A question mark in parentheses (?) denotes that no relevance score was given.

#	Topic	Result and Comment
1	what symbol is ζ	Riemann zeta function(4) other results may exist
2	define $\iff$ in $A \iff B$	Logical equivalence(4), If and only if(?)
3	definition $a \oplus b$	Direct sum(4), $\oplus$(?)
4	$\begin{vmatrix} a & b \\ c & d \end{vmatrix}$	Determinant$^+$(4)
5	define notation ${}_2F_1(a, b; c; z)$	Hypergeometric function$^+$(4)
6	$a_0 + \frac{b_1}{a_1+} \frac{b_2}{a_2+} \cdots$ define pattern	Continued fraction$^+$(4)
7*	$PY(*1*)$	Pitman-Yor process(?), Classical general equilibrium model$^+$(0)
8	$n = \frac{c}{*1*}$ light	Refractive index(4) $*1*$ denotes the frequency ν.
9	$*1*_{n+1} = r*1*_n(1 - *1*_n)$ recurrence relation	Bifurcation diagram$^+$(4)
10	$g(x) = \frac{1}{1+e^{-x}}$	Logistic function(4), Multimodal learning(0)
11	$F = ma$	Newton's laws of motion$^+$(2) Comment: The authors consider Newton's second law as a relevant result for the query $F = ma$., Force(4)
12	Legendre $\left(\frac{a}{p}\right)$	Legendre symbol(4)
13	find $ax^2 + bx + c = (*1*)^2 + *2*$	Quadratic equation(4), Quadratic function(2), Binomial theorem(?)
14*	convert $log_2(*1*)$ to $\ln(*1*)$	Binary logarithm$^+$(3)
15	compute value for $\binom{n}{k}$	Binomial coefficient(4)
16	solve $x_n^2 - x_{n-1}x_{n+1} = c$ for x_n	No result! At first glance, $x_n = \sqrt{c - x_{n-1}x_{n+1}}$ appears to be related to the Mandelbrot set(?), but the $n + 1$ index is hard to interpret.
17	factor $x^3 + Dy^3 + D^2z^3 - 3Dxyz$ multiple variables	Hessian form of an elliptic curve(2)
18	$\lim_{x\to 0} \frac{2 - \cos(3x) - \cos(4x)}{x}$ solve limit	L'Hôpital's rule(3)
19	sequence name $1, 2, 2, 3, 3, 4, 4, 4, 5, 5, \ldots$	Golomb sequence(4)
20	why $*1*^2 - 7*1* + 2$ polynomial but $\frac{*1*^2-7*1*+2}{*1*+2}$ not polynomial	Polynomial(4)
21	difference between Log $*1*$ and log $*1*$	Common logarithm(0) Comment: This article elaborates on the difference between *Log* and log in detail.
22*	explanation intuition $\nabla \times E = -\frac{\delta B}{\delta t}$	Faraday's law of induction(2) Comment: The article explains the Maxwell–Faraday equation $\nabla \times E = -\frac{\partial B}{\partial t}$ in detail
23*	is $P = NP$ possible	P versus NP problem(4)
24*	what is gamma $\int_0^{*1*} e^{-x}dx$	Gamma function(4)
25	prove $(f \circ g)' = (f' \circ g) \cdot g'$	Chain rule(4)
26	prove $x^2 + y^2 = z^2$	Pythagorean theorem(4)
27	$*1*^2 + *2*^2 = *3*^2$ uses	Pythagorean triple(3), Polar coordinate system(3), Cauchy-Schwarz inequality(3), Euclidean vector$^+$(2), Parallelogram law(2)
28	example uses $f(a_1x_1 + \cdots + a_nx_n) \leq a_1f(x_1) + \cdots + a_nf(x_n)$ where $x_i \in \mathbb{R}$ $a_i \geq 0$ $\sum_{i=1}^n a_i = 1$	Sublinear function(1), Subadditivity$^+$(1)
29*	application growth $\frac{dP}{dt} = r\left(1 - \frac{P}{K}\right)P$ epidemiology biology	Logistic function$^+$(4), R/K selection theory(4)
30	applications $f \star g$ where $(f \star g)[*1*] := \sum_{*2*=-\infty}^{\infty} f(*2*)\, g(*1* - *2*)$	Convolution$^+$(2) Comment: The formula in the query is neither an exact match for convolution $(f * g)[*1*] := \sum_{*2*=-\infty}^{\infty} f[*2*]\, g[*1* + *2*]$ nor for cross correlation(4) $(f \star g)[*1*] := \sum_{*2*=-\infty}^{\infty} f^*[*2*]\, g[*1* - *2*]$. Note that for f non Hermitian $f \star g = f^*(-t) * g$.

*) Note that the incorrect typesetting was given in the topics.
+) The link points to a specific section of the Wikipedia article.

Table 5: Compiled gold standard: Our original results are stated in boldface in case we still consider them as relevant, or stroked through in case we reconsidered our decision given the assessors' feedback. Normal font indicates results of other participants that we included in the gold standard.

#	Topic	Result and Comment
1	what symbol is ζ	**Riemann zeta function**, Damping ratio, Hurwitz zeta function, 1s Slater-type function, Jerk (physics), Oblate spheroidal coordinates, Routhian mechanics Comment: 6 new hits.
2	define $\iff$ in $A \iff B$	Monoidal t-norm logic, **Logical equivalence**, **If and only if**, Logical biconditional Contraposition, (Tautology (logic) or Exportation (logic)) Comment: 4 new hits; the assessors rated 1 hit as irrelevant that we consider relevant.
3	definition $a \oplus b$	**Direct sum**, $\oplus$, Exclusive or Comment: 1 new hit.
4	$\begin{vmatrix} a & b \\ c & d \end{vmatrix}$	**Determinant**, Laplace expansion Comment: 1 new hit.
5	define notation ${}_2F_1(a, b; c; z)$	**Hypergeometric function** Comment: no new hits.
6	$a_0 + \frac{b_1}{a_1+} \frac{b_2}{a_2+} \cdots$ define pattern	Generalized continued fraction, **Continued fraction** Comment: 1 new hit.
8	$n = \frac{c}{*1*}$ light	**Refractive index** Comment: no new hits.
9	$*1*_{n+1} = r*1*_n(1 - *1*_n)$ recurrence relation	Logistic map **Bifurcation diagram**, Lyapunov fractal Comment: 2 new hits (one more and one less relevant than our original hit).
10	$g(x) = \frac{1}{1+e^{-x}}$	**Logistic function**, ~~Multimodal learning~~ Comment: No new hits; we consider the second hit as a duplicate.
11	$F = ma$	**Newton's laws of motion**, **Force** Comment: no new hits.
12	Legendre $\left(\frac{a}{p}\right)$	**Legendre symbol** Comment: no new hits.
13	find $ax^2 + bx + c = (*1*)^2 + *2*$	**Quadratic equation**, **Quadratic function**, **Binomial theorem** Comment: no new hits.
14	convert $log_2(*1*)$ to $\ln(*1*)$	**Binary logarithm** Comment: no new hits.
15	compute value for $\binom{n}{k}$	**Binomial coefficient** Comment: no new hits.
18	$\lim_{x\to 0} \frac{2 - \cos(3x) - \cos(4x)}{x}$ solve limit	**L'Hôpital's rule**, List of limits Comment: 1 new hit.
19	sequence name $1, 2, 2, 3, 3, 4, 4, 4, 5, 5, \ldots$	**Golomb sequence** Comment: no new hits.
20	why $*1*^2 - 7*1* + 2$ polynomial but $\frac{*1*^2-7*1*+2}{*1*+2}$ not polynomial	**Polynomial** Comment: no new hits.
21	difference between Log $*1*$ and $\log *1*$	**Common logarithm** Comment: no new hits.
22	explanation intuition $\nabla \times E = -\frac{\delta B}{\delta t}$	**Faraday's law of induction** Comment: no new hits.
23	is $P = NP$ possible	**P versus NP problem** Comment: no new hits.
24	what is gamma $\int_0^{*1*} e^{-x} dx$	**Gamma function**, Incomplete gamma function Comment: 1 new hit.
25	prove $(f \circ g)' = (f' \circ g) \cdot g'$	**Chain rule** Comment: no new hits.
26	prove $x^2 + y^2 = z^2$	**Pythagorean theorem** Comment: no new hits.
27	$*1*^2 + *2*^2 = *3*^2$ uses	**Pythagorean triple**, **Polar coordinate system**, **Cauchy-Schwarz inequality**, **Euclidean vector**, **Parallelogram law**, Crossed ladders problem, Cauchy–Schwarz inequality, Law of cosines, Isosceles triangle, Slant height, Special right triangles, Triangle Comment: 7 new hits (no particular order of relevance).
28	example uses $f(a_1x_1 + \cdots + a_nx_n) \leq a_1f(x_1) + \cdots + a_nf(x_n)$ where $x_i \in \mathbb{R}$ $a_i \geq 0$ $\sum_{i=1}^{n} a_i = 1$	**Jensen's inequality**, ~~Sublinear function~~, ~~Subadditivity~~ Comment: 1 new hit that we consider significantly more relevant than our hits.
29	application growth $\frac{dP}{dt} = r\left(1 - \frac{P}{K}\right)P$ epidemiology biology	**Logistic function**, **R/K selection theory**, Ecology Comment: 1 new hit.
30	applications $f \star g$ where $(f \star g)[*1*] := \sum_{*2*=-\infty}^{\infty} f(*2*)\, g(*1* - *2*)$	**Convolution**, Cross-correlation Comment: 1 new hit; both hits are equally relevant and very similar.