DESIGN AND IMPLEMENTATION OF A QUESTION-ANSWERING SYSTEM FOR MATH-BASED QUERY

Arnab Paik(1181200069)
Brajanandan Gupta(1181200070)
Uddeshya Raj(1181200071)

Under the guidance of
Dr. Sourish Dhar

Dept. of CSE
Assam University, Silchar

August 11, 2023

CONTENTS

Introduction

What is Question-Answering System

- Question answering (QA) is a computer science discipline within the fields of information retrieval and natural language processing (NLP), which is concerned with building systems that automatically answer questions posed by humans in a natural language.

  .
- Question answering research attempts to deal with a wide range of question types including: fact, list, definition, How, Why, hypothetical, semantically constrained, and cross-lingual questions.

MOTIVATION

- Math has always been a tough subject and help is not always available, especially high school and university level math.
- NLP and Question Answering are already new fields and math specific question answering is still newer.
- Recognition of math based terms and/or formulae by existing search engines may provide better results on certain queries.
- Creation of math specific search engines which can provide relevant topics from internet given a question (like Google Scholar focuses specifically on scholarly articles).
- Quickly able to retrieve related information from a formula. Can be useful for students in remote areas who lack proper teachers and resources.
- Can be used to make a Wolfram|Alpha like platform which can be used to solve math questions but more powerful.

Problem Description

ARQMath-3

- **Given a posted question as a query, search all the answer posts and return relevant answer posts.**
- Given a question post with an identified formula as a query, search all question and answer posts and return relevant formulas with their posts.
- Given a posted question as a query, return a single answer. The answer may be automatically generated, and may contain passages from outside the ARQMath collection.

.

ANSWER RETRIEVAL IN ARQMATH

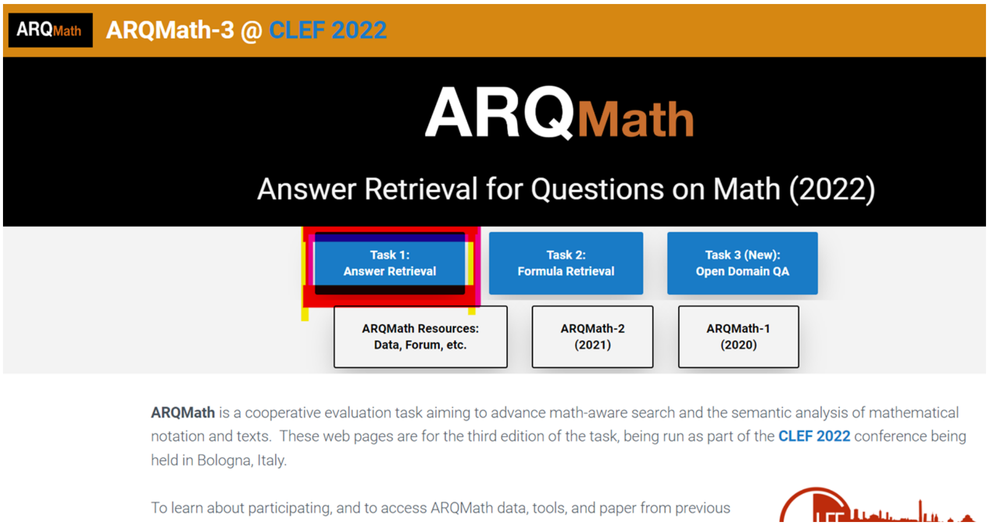


Figure. 1

AN EXAMPLE OF ANSWER RETRIEVAL

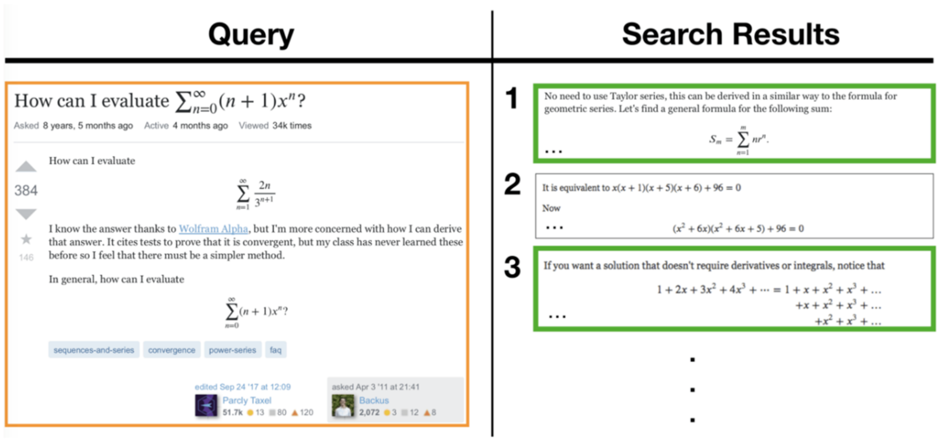


Figure. 1.1

OBJECTIVES OF THE PROJECT

- Develop a Mathematics Question Answering system that utilizes NLP techniques to process and understand a variety of mathematical questions..
- Explore and implement mathematical language models or representations that aid in accurately interpreting user queries..
- Design an efficient answer retrieval mechanism that can search through mathematical literature and resources to locate and return a list of relevant answers.
- Evaluate the developed system's performance through benchmark datasets, comparing its answers with expert-validated responses.

CHALLENGES

- Unavailability of proper systems that understand Mathematical semantics.
- Limited research done in this domain.
- Capturing and utilizing context effectively to provide contextually relevant answers.
- Mathematical Expression Similarities.
- Difficulty in representing formulae in proper data-structure that it is easy to parse.

DATASET


```
<row Id="5" PostTypeId="1" CreationDate="2010-07-20T19:18:01.250" ViewCount="12722" Score="54"
CommentCount="7" OwnerUserId="38" Title="How can you prove that the square root of two is
irrational?" Body="&lt;p&gt;I have read a few proofs that &lt;span class=&quot;math-container&quot;
id=&quot;39&quot;&gt;\sqrt{2}&lt;/span&gt; is irrational.&lt;/p&gt;  &lt;p&gt;I have never,
however, been able to really grasp what they were talking about.&lt;/p&gt;  &lt;p&gt;Is there a
simplified proof that &lt;span class=&quot;math-container&quot;
id=&quot;40&quot;&gt;\sqrt{2}&lt;/span&gt; is irrational?&lt;/p&gt; " AnswerCount="14"
AcceptedAnswerId="7" Tags="&lt;elementary-number-theory&gt;&lt;proof-
writing&gt;&lt;radicals&gt;&lt;rationality-testing&gt;"/>
```


Figure. A post in the dataset


```
{'Id': '5', 'PostTypeId': '1', 'CreationDate': '2010-07-20T19:18:01.250', 'ViewCount': '12722', 'Score': '54', 'CommentCount': '7', 'Ow
nerUserId': '38', 'Title': 'How can you prove that the square root of two is irrational?', 'Body': '<p>I have read a few proofs that <s
pan class="math-container" id="39">\\sqrt{2}</span> is irrational.</p>  <p>I have never, however, been able to really grasp what they w
ere talking about.</p>  <p>Is there a simplified proof that <span class="math-container" id="40">\\sqrt{2}</span> is irrational?</p> ',
 'AnswerCount': '14', 'AcceptedAnswerId': '7', 'Tags': '<elementary-number-theory><proof-writing><radicals><rationality-testing>'}
```


Figure. Parsed post from dataset

Attributes contained in posts.xml:

- id
- PostTypeId -> 1: Question, 2: Answer
- ParentID (only present if PostTypeId is 2)
- AcceptedAnswerId (only present if PostTypeId is 1)
- Score
- Title
- Tags
- Body
- AnswerCount
- CommentCount
- ViewCount
- CreationDate
- OwnerUserid
- LastEditorUserid
- LastEditDate
- LastEditorDisplayName
- LastActivityDate
- CommunityOwnedDate
- ClosedDate
- FavouriteCount

LITERATURE SURVEY

SI No.	Paper Name	Author(s)	Published at (year)	Contribution	Limitations .
1	DPRL Systems in the CLEF 2020 ARQMath Lab	Behrooz Mansouri, Douglas W. Oard and Richard Zanibbi	CLEF 2020, Thessaloniki, Greece (2020)	Usage of Tangent-CFT to embed mathematical formulas and parse them and use Re-Ranking to increase relevancy of result.	Accuracy is slightly lower than base line system in nDCG evaluation measure. Re-ranking significantly increases the retrieval time.
2	Dowsing for answers to math questions: Doing better with less	Andrew Kane, Yin Ki Ng and Frank Wm. Tompa	CLEF 2022: Conference and labs of the evaluation forum (2022)	Indexing and query principles that they describe can be used with any search engine to make it math-aware	Query execution efficiency is low.

LITERATURE SURVEY

SI No.	Paper Name	Author(s)	Published at (year)	Contribution	Limitations .
3	Combined sparse and dense information retrieval	Vit Novotny and Michal Stefanik	CLEF 2022: Conference and labs of the evaluation forum (2022)	Usage of soft vector space model improves effectiveness of the model compared to sparse models	Soft vector space model does not fully exploit the semantic information given in the source. Loss of ability to model the similarity between text and math tokens.
4	Transformer-Encoder and Decoder Models for Questions on Math	Anja Reusch, Maik Thiele and Wolfgang Lehnar	CLEF 2022: Conference and labs of the evaluation forum (2022)	Usage of transformer models like BERT, RoBERTa and ALBERT for retrieval of answers given a mathematical question	Training such models on large datasets is resource intensive and requires very high end GPUs

LITERATURE SURVEY

SI No.	Paper Name	Author(s)	Published at (year)	Contribution	Limitations .
5	DPRL Systems in CLEF 2022: Introducing MathAMR for Math-Aware Search	Behrooz Mansouri, Douglas W. Oard and Richard Zanibbi	CLEF 2022: Conference and labs of the evaluation forum (2022)	Introduced a abstract meaning representation system for maths (MathAMR)	Low accuracy of the new MathAMR. 50 persent accuracy in best case scenario.
6	Approach Zero and Anserini at CLEF-2021 ARQMath Track: Applying Substructure Search and BM25 on Operator Tree Path	Wei Zhong, Xinyu Zhang, Ji Xin, Richard Zanibbi and Jimmy Lin	CLEF 2021: Conference and Labs of the Evaluation Forum (2021)	Use of Approach Zero, a structure aware search system, and Anserini, full-text retrieval system to solve task-1.	Task-1 results not very competitive. Established that text-only retrieval systems perform better than Approach Zero.

LITERATURE SURVEY

SI No.	Paper Name	Author(s)	Published at (year)	Contribution	Limitations .
7	Information Retrieval Based on Stochastic Models	Masaki Murata, Kiyotaka Uchimoto, Hiromi Ozaku, Hitoshi Isahara	NTCIR 1: Communications Research Laboratory, Ministry of Posts and Telecommunications, Japan	Provided pointers on implementation of text-only retrieval models	This paper talks about document retrieval in general / math specific may be needed. No info on how to implement dictionary for query expansion or how query expansion actually works.
8	Proposal and Evaluation of Significant Words Selection Method based on AIC	Shigeki Ohira, Katasuhiko Shirai	NTCIR 1: School of Science and Engineering, Waseda University, Tokyo	Usage of Chi-Square method and AIC (Akaike's Informationtheoretic Criterion) for selecting significant words.	Paper was written for adhoc Japanese question answering system. No info about actually implementing the system.

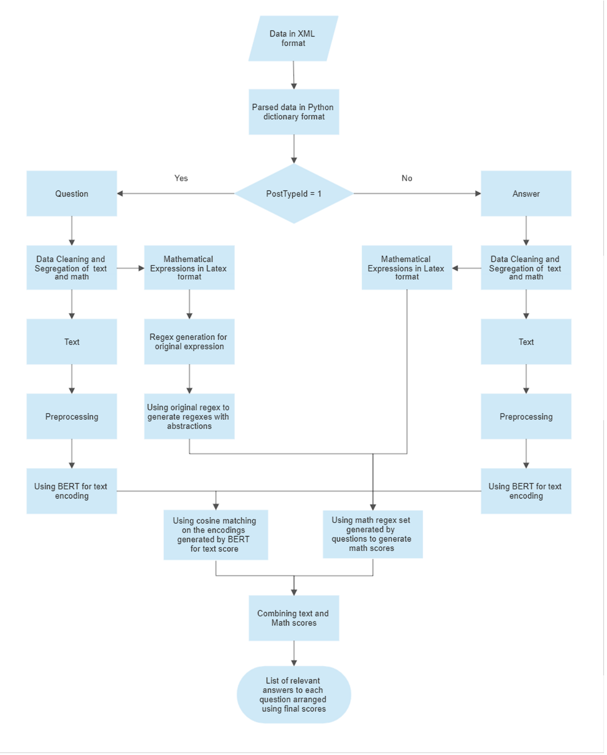
LITERATURE SURVEY

Summary of Literature Survey

- Lack of any concrete math-aware system is the cause of generally low accuracy of these systems.
- Each submission with competitive accuracy required extremely high end systems to train and implement models. Especially systems using transformer models.
- Symbol layout tree is the main way to process math expressions in these IR models.
- Some filtering will be necessary to trade-off accuracy for available resource.

Methodology

Proposed high level workflow



Methodology

Procedure for text searching

Text matching is straightforward using standard procedure for these type of tasks.

- **Step 1:**Separation of text from math
- **Step 2:**Removal of XML tags and symbols
- **Step 3:**Stemming and tokenization of words
- **Step 4:**Creating a bag of words using those processed words
- **Step 5:**Generating Text Encoding using BERT model
- **Step 6:**Using cosine similarity to rank posts according to match score generated

METHODOLOGY
WHY USE BERT?

- This model generates embeddings, creating a matrix with 768 features for each post.
- The embeddings consider semantic meaning and capture contextual word relationships.
- BERT is robust to noise due to training on vast text data, unlike TF-IDF affected by errors.
- TF-IDF's strength lies in specific tasks, while BERT's versatility results from general language pattern learning
- BERT stands out as a powerful and adaptable embedding technique, making it the prime choice for diverse NLP tasks

Methodology

Problems with mathematical expression

- Multiple levels of nesting and abstraction make exact expression matching futile.
- Multiple notations exist for same expression. For example: nCr is also written as C(n,r) or nCr . Normalization of mathematical notations is required.
- Greek letters and numbers create ambiguity about are they to be considered for exact match or to be replaced with wildcard. Semantics/context of their use matters in such situation.
- Matrix and determinants are tough to deal with in any scenario.

METHODOLOGY

REGULAR EXPRESSION CAN HELP

- Regular expressions are patterns used to match character combinations in a string.
- Math expressions are written using Latex in the posts which can easily be read and worked on as strings. Unlike MathML having the whole expression in a single line, without any hierarchy to be parsed, also helps.
- Regular expressions are very fast so no need to use slower and more complex sub/string matching algorithms.
- Symbol Layout trees use n-gram search which may help in finding similar looking expression with partial matches but properly generated RegEx can do it much effectively in a way which is more semantically correct.

METHODOLOGY

MAIN REGEX GENERATION

- Take an expression $a^2 + b^2 = c^2$

- In LaTeX it will be written as $a\hat{}\{2\} + b\hat{}\{2\} = c\hat{}\{2\}$

- For given expression multiple regex can be generated with multiple levels of abstraction. For example to take into account the variations in variable name only, regex r'$([A-Za-z]+)\hat{}\{2\} + ([A-Za-z]+)\hat{}\{2\} = ([A-Za-z]+)\hat{}\{2\}'$ can be generated, where **[A-Za-z]+** represents any English string that can be used as variable name. Including Greek letters into it is also easy with some slight modifications to the regex.
- Similarly regex **r'[A-Za-z]+(**$\hat{}\{-?[0-9]+\})?+|-)*[A-Za-z]+(\hat{}\{-?[0-9]+\})?'$ can match with Latex representation of any linear Algebraic expression without any constant. Lets assign it to a variable A.

Methodology

Main Regex generation

- Consider another regex **r’(+|-)?[0-9]+’** which will match all numeric constants and store it in variable B.
- Combining A and B in a regex **‘(?<!/w)A?(?:B)?(?!/w)(?<!/W(?!/w))(?<!**$^{(}?!/w))'$
  will make it able to match all linear algebraic expressions. This can further be used to create regex for more complex math expressions using nesting. Here /w is a word type character(a-z, A-Z, 0-9,) and /W is a non-word type character.

Methodology

Partial Regex generation

```python
og = processTagsAndVariables(math_exp)
all_regex = set()
all_regex.add(og)
create_all_regexes(og, all_regex)
for regex in all_regex:
    print(regex)
```

```
[a-zA-Z]+\\frac\{\\tanh\{.+?}}\{.+?}
[a-zA-Z]+\\frac\{.+?}\{\\log_2\{.+?}}
[a-zA-Z]+\\frac\{.+?}\{\\log_2\{[a-zA-Z]+}}
[a-zA-Z]+\\frac\{\\tanh\{(((\+|-)?([0-9]+)(\.[0-9]+)?)|((\+|-)?\.?[0-9]+))}}\{\\log_2\{.+?}}
[a-zA-Z]+\\frac\{.+?}\{.+?}
[a-zA-Z]+\\frac\{\\tanh\{.+?}}\{\\log_2\{.+?}}
[a-zA-Z]+\\frac\{\\tanh\{(((\+|-)?([0-9]+)(\.[0-9]+)?)|((\+|-)?\.?[0-9]+))}}\{.+?}
[a-zA-Z]+\\frac\{\\tanh\{.+?}}\{\\log_2\{[a-zA-Z]+}}
[a-zA-Z]+\\frac\{\\tanh\{(((\+|-)?([0-9]+)(\.[0-9]+)?)|((\+|-)?\.?[0-9]+))}}\{\\log_2\{[a-zA-Z]+}}
```

Figure. Partial Regex Generated

METHODOLOGY

DOWNSIDE OF USING REGEX

- Searching using regex provides no score so two matching formulae/expressions can't be compared with each other. Another way of score generation needs to be implemented.
- If query is a simple expression then it is possible to find a match in a complex expression using nesting of expressions, but if the expression in query is more complex than the one available in answer posts matching may not work. For example: a^2 as query will match with $tan^2\theta$, but not the other way around. Though another solution exists for this.

Methodology

Score Generation from regex matching

- Regex can provide only True and False results (formula matches or not)
- To generate Floating point scores for math expression matching multiple regular expression are generated from the main expression with multiple levels of abstraction.
- The fraction of all the regular expression generated that actually find a match is used as the score for expression matching.

RESULTS

PRECISION

	top5	top10	top15	top20	top25	top30
ques 1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques 2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques 3	1.000000	0.800000	0.533333	0.400000	0.320000	0.266667
ques 4	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques 5	0.200000	0.100000	0.066667	0.050000	0.040000	0.033333
ques 6	0.200000	0.100000	0.066667	0.050000	0.040000	0.033333
ques 1099	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques 1100	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques 1101	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques 1102	0.800000	0.400000	0.266667	0.200000	0.160000	0.133333
ques 1103	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ques 1104	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
Average	0.782609	0.731431	0.701087	0.679393	0.662609	0.649758

Figure. Precision for top results

Results

Recall

	top5	top10	top15	top20	top25	top30
ques1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ques2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques3	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques4	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques5	1.000000	1.000000	0.500000	0.500000	0.500000	0.500000
ques1101	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ques1102	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
ques1103	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
ques1104	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Average	0.461730	0.516606	0.543327	0.564463	0.582397	0.592588

Figure. Recall for top results

Results

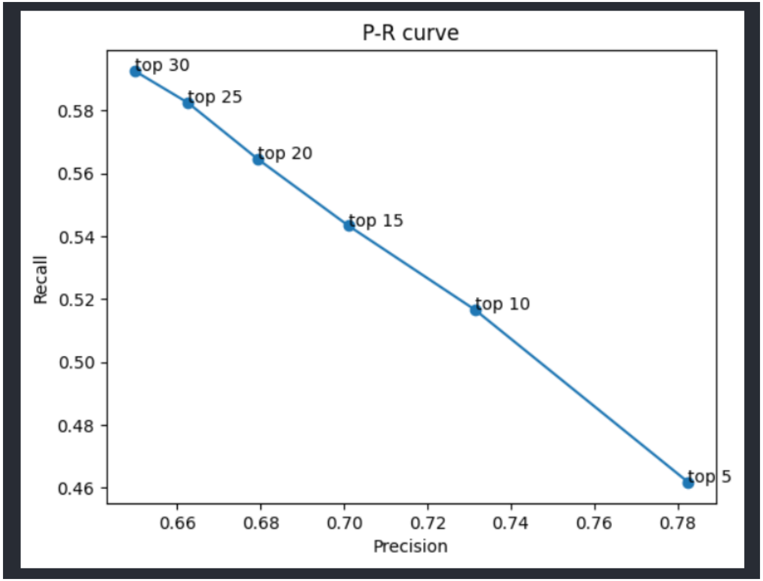


Figure. P-R curve

Results

Analysis

- In the P-R curve the precision decreases and recall increases as we take into consideration more and more posts
- Since the posts are sorted in descending order in terms of relevancy, as we go down and include more posts, the count of irrelevant posts increases and precision decreases.
- As we increase the no. of posts to be counted in final result from top 5 to top 10 to top 30, the relevant results which were not earlier included also gets counted, thus increasing the recall.
- Precision and recall are inversely correlated

FUTURE WORK

- Increase the speed of model using GPU acceleration in tensor-flow.
- Enhance the accuracy using better data cleaning techniques.

References

- 1. Andrew Kane, Yin Ki Ng, and Frank Wm. Tompa. Dowsing for answers to math questions: Doing better with less. In Proceedings of the 2022 Conference and Labs of the Evaluation Forum (CLEF), pages 157–170. CEUR-WS.org, 2022.
- 2. Zhe Liu, Yu Wang, Rui Chen, Zhe Lin, and Maosong Sun. Bert-based text matching for question answering. arXiv preprint arXiv:1901.07888, 2019.
- 3. Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bertnetworks. arXiv preprint arXiv:1908.10084, 2019.
- 4. Anja Reusch, Maik Thiele, and Wolfgang Lehnar. Transformer-encoder and decoder models for questions on math. In Proceedings of the 2022 Conference and Labs of the Evaluation Forum (CLEF), pages 171–184. CEUR-WS.org, 2022.
- 5. Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Combined sparse and dense information retrieval. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 582–592. Association for Computational Linguistics, 2020.
- 6. Cathy O'Neil. Data Cleaning: A Practical Introduction. O'Reilly Media, 2013

Thank You