The MCAT Math Retrieval System for NTCIR-11 Math Track

Giovanni Yoko Kristianto Goran Topic Florence Ho Akiko Aizawa

National Institute of Informatics

MCAT Math Retrieval

- Objective:
  - Provide a search system for math expressions
- Approach:
  - Encode structure and tokens of each expression
    - For ranking, structure has higher priority than tokens of expression
  - Enable full-text search and math browsing system
    - Extract textual information for each expression

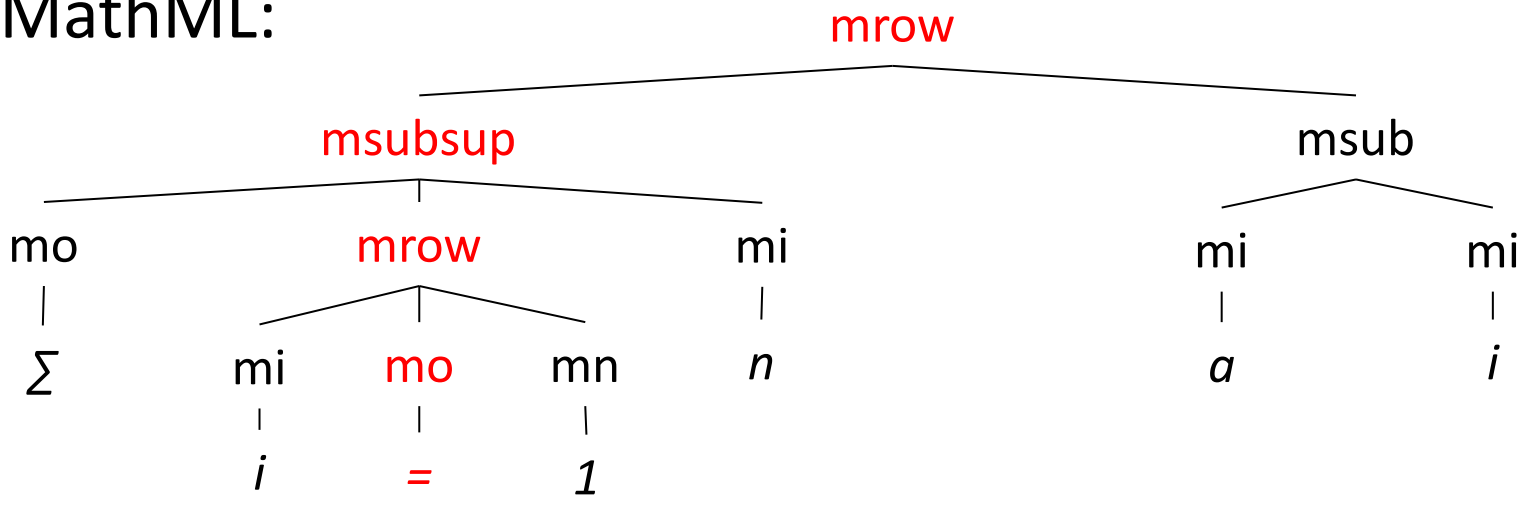
Overview

- Database:
  - Apache Solr (Lucene)
- Indexed data:
  - Structure and tokens of each math expression
  - Textual information of each math expression
- Post-retrieval reranking method

Indexing of Math Expressions

$$\left(\sum_{i=1}^{n} a_i x^i\right)$$

- MathML:



- Encoding results
  - opaths: 1#2#2#mo#=
  - upaths: ###mo#=
  - sisters: mi#i mo#= mn#1

Textual Information

MATH Context

The notation *T* refers to a set of topics, *V* to the word vocabulary and $g_i(\alpha_i)$

to the Dirichlet distribution associated with topic t_i.

MATH Description

The notation *T* refers to a set of topics, *V* to the word vocabulary and $g_i(\alpha_i)$

to the Dirichlet distribution associated with topic t_i.

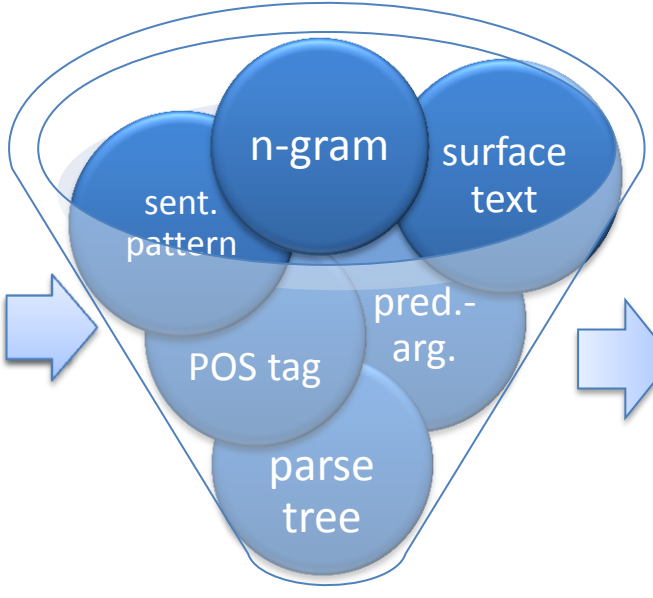
Description Extraction

Noun Phrases
as description candidates

(Math, NP)

- (Math_1, NP_1)
- (Math_1, …)
- (Math_1, NP_n)
- ….
- (Math_n, NP_1)
- (Math_n, …)
- (Math_n, NP_n)

Features Extraction



Description Extraction Results

$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = N(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1})$
likelihood function for least square

$N(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1})$
normal distribution

$\boldsymbol{x}$
input variables

β
precision

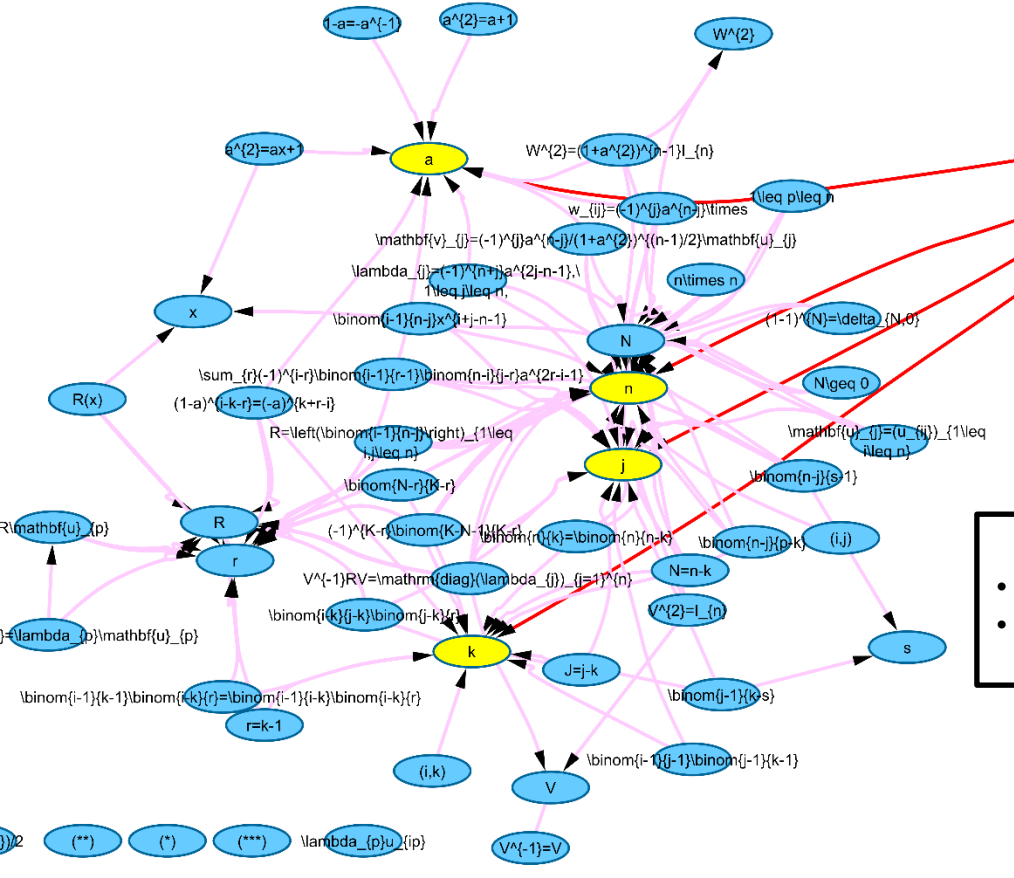
$\boldsymbol{w}$
parameter

Training set: NTCIR-10 Math Understanding Subtask dataset

Extraction on NTCIR-11 dataset: 360 hours (approx.)

Dependency Graph

Query: $x^2 - x - 1 = 0$ *(golden ratio; Fibonacci)*



$$u_{ij} = \sum_{k=1}^{j} (-1)^{i-k} \binom{i-1}{k-1} \binom{n-i}{j-k} a^{2k-i-1}$$

n	k	j	a
• rows • the upper parameter			• the **golden ratio** • a root

Post-Retrieval Reranking

- Given our database schema, we cannot force Lucene to emphasize content or emphasize structure for ranking.

- Given a query formula f_q and its retrieved formula f_x, calculate $formSim$:

  $$formSim(f_x, f_q) = \lambda \cdot contSim(f_x, f_q) + (1 - \lambda) \cdot structSim(f_x, f_q)$$

  λ : weight ([0..1])
  $contSim$: content-based similarity (Euclidean)
  $structSim$: path-based structure similarity (Euclidean)

- $finalScore(f_x, f_q) = formSim(f_x, f_q) \cdot luceneScore(f_x, f_q)$

Ranking Performance

Runs	High Relevancy			Partial Relevancy		
	MAP	P@5	P@10	MAP	P@5	P@10
Context	.0515	.1640	.0960	.0776	.4040	.2400
Description	.0648 ($1.4*10^{-2}$)	.1920 ($3.2*10^{-2}$)	.1160 ($4.1*10^{-3}$)	.0860 ($4.6*10^{-2}$)	.4160 ($2.3*10^{-1}$)	.2600 ($2.2*10^{-2}$)
Context+ Description	.0718 ($1.1*10^{-3}$)	.2120 ($1.8*10^{-4}$)	.1240 ($5.6*10^{-4}$)	.0926 ($3.0*10^{-3}$)	.4480 ($5.8*10^{-3}$)	.2800 ($1.5*10^{-4}$)
Description+ Rerank	.0742 ($6.5*10^{-3}$)	.2080 ($1.4*10^{-1}$)	.1300 ($3.2*10^{-2}$)	.0933 ($4.7*10^{-2}$)	.4640 ($1.7*10^{-2}$)	.2800 ($4.5*10^{-2}$)

** (…) represents the p-value*

System Performance

- Indexing time:
  - Encoding formulae : 40 hours
  - Importing data into Solr : 8 hours
- Solr size (disk): 58Gb
- Tomcat for Solr: 300Mb

Conclusion

- Descriptions extraction, dependency graph, and post-retrieval reranking methods effectively improved the ranking performances.

- If $c = 0$, then Alice sends the element t to Bob, and Bob checks if the equality $v = t(u)$ is satisfied. If it is, then Bob accepts the authentication.
- If $c = 1$, then Alice sends the composition ts to Bob, and Bob checks if the equality $v = ts(x)$ is satisfied. If it is, then Bob accepts the authentication.

2.2 Protocol II

In this protocol, the hardness of obtaining the "permanent" private key for the adversary can be based on "most any" search problem; we give some concrete examples i
give a generic protocol.

1. Alice's public key consists of a set S that has a property $\mathcal{P}$. Her private key is a proof (or a "witness") that S does have this property. We are also assuming that the
2. To begin authentication, Alice selects an isomorphism $\varphi: S \to S_1$ and sends the set S_1 (the commitment) to Bob.

$\varphi: S \to S_1$

an isomorphism

Components

φ an endomorphism such that $\varphi(g) = h$
S a set that has a property $\mathcal{P}$
S_1 the commitment

Similar expressions

$\varphi': S \to S_1'$	an isomorphism
$g_t: S \to S^1$	a quotient map obtained by projecting the bicollar on each $c_i(t)$ to an interval of length $2\pi t_i$ and collapsing each complementary cobordism $C_i(t)$ to a
$g_t: S \to S^1$	a family
$\varphi: \Gamma \to \Gamma_1$	an isomorphism
$\varphi: \Gamma \to \Gamma_1$	an isomorphism
$\varphi: D^{n+1} \to S^{n+2}$	