The MCAT Math Retrieval System for NTCIR-10 Math Track

Goran Topić Giovanni Yoko Kristianto Minh-Quoc Nghiem
Akiko Aizawa

National Institute of Informatics

2013

Introduction

Introduction

Unneccessary

Math Understanding

Objective

Extract descriptions of math expressions from the text

Math Understanding

Objective

Extract descriptions of math expressions from the text

Assumption

Descriptions are noun phrases in the same sentence as the math expression

Math Understanding

Objective

Extract descriptions of math expressions from the text

Assumption

Descriptions are noun phrases in the same sentence as the math expression

Method

Identify noun phrases using Stanford parser
Train a linear-kernel SVM classifier for (math, description) pairs

SVM Features

Apposition?	set $\mathbb{N}$
Colon?	set: $\mathbb{N}$
Comma?	set, $\mathbb{N}$
Intervening expression?	set . . . A . . . $\mathbb{N}$
Parenthetical?	$\mathbb{N}$ (set)
Word distance	set *(4 words)* $\mathbb{N}$
After description?	set $\mathbb{N}$
2-word description context	in/IN the/DT set/NN $\mathbb{N}$/MATH of/IN
3-word expression context	in/IN the/DT set/NN $\mathbb{N}$/MATH of/IN natural/JJ numbers/NNS
First word of description	set/NN
Last word of description	set/NN
Unigrams	in/IN, the/DT, set/NN, $\mathbb{N}$/MATH, of/IN, natural/JJ, numbers/NNS
Bigrams	in/IN the/DT, the/DT set/NN, set/NN $\mathbb{N}$/MATH, $\mathbb{N}$/MATH of/IN
Trigrams	in/IN the/DT set/NN, set/NN $\mathbb{N}$/MATH of/IN
First intervening verb	set . . . shows . . . $\mathbb{N}$

Experiment

Baseline

Noun phrase in apposition to the mathematical expression

Experiment

Baseline

Noun phrase in apposition to the mathematical expression

Runs

Full descriptions, all features
Full descriptions, without apposition feature
Short descriptions, all features
Short descriptions, without apposition feature

Results

Run	P	R	F-1
Strict Matching Evaluation			
full, baseline	43.10	23.85	30.96
MCAT: full, all features	61.94	37.03	46.35
MCAT: full, no apposition	61.92	37.33	46.58
short, baseline	55.35	29.94	38.86
MCAT: short, all features	68.24	40.42	50.77
MCAT: short, no apposition	67.67	40.22	50.45
Soft Matching Evaluation			
full, baseline	64.21	34.73	45.08
MCAT: full, all features	85.48	47.41	61.24
MCAT: full, no apposition	87.25	48.30	62.18
short, baseline	64.21	34.73	45.08
MCAT: short, all features	81.68	42.81	56.18
MCAT: short, no apposition	81.24	42.61	55.90

Math Understanding Conclusions

Conclusions

Apposition feature slightly harms full description extraction (because of discontinuous descriptions), but helps short description extraction (since they mostly are appositions)

Unsurprisingly, soft matching works better on full descriptions than on short descriptions (larger text region: more chances of overlap)

In strict matching runs, 64% of full and 75% of short descriptions appear as noun phrases.

In soft matching runs, 89% of full and 87% of short descriptions appear as noun phrases.

Math Retrieval

Objective

Provide a search system for mathematical expressions

Math Retrieval

Objective

Provide a search system for mathematical expressions

Requirements

Flexibility: focus on recall (sacrifice precision)
Encode structure as well as tokens
Allow full-text search on descriptions

Math Retrieval

Objective

Provide a search system for mathematical expressions

Requirements

Flexibility: focus on recall (sacrifice precision)

Encode structure as well as tokens

Allow full-text search on descriptions

Engine

Starting with full-text search requirement, we selected Apache Solr (Lucene) for our database.

Indexing

Papers

Papers are placed in a separate index, with full-text, language-dependent fields for titles and abstracts, and general fields for authors.

Indexing

Papers

Papers are placed in a separate index, with full-text, language-dependent fields for titles and abstracts, and general fields for authors.

Expressions

Expression descriptions are indexed as full-text, language-dependent fields. Expression MathML structure is encoded into three whitespace-separated, multivalued fields: ordered paths, unordered paths and sisters.

Indexing

Papers

Papers are placed in a separate index, with full-text, language-dependent fields for titles and abstracts, and general fields for authors.

Expressions

Expression descriptions are indexed as full-text, language-dependent fields. Expression MathML structure is encoded into three whitespace-separated, multivalued fields: ordered paths, unordered paths and sisters.

A Note on Descriptions

Due to time constraints, we did not use the extracted descriptions, but used a fixed 10-word context instead.

Indexing

Papers

Papers are placed in a separate index, with full-text, language-dependent fields for titles and abstracts, and general fields for authors.

Expressions

Expression descriptions are indexed as full-text, language-dependent fields. Expression MathML structure is encoded into three whitespace-separated, multivalued fields: ordered paths, unordered paths and sisters.

A Note on Descriptions

Due to time constraints, we did not use the extracted descriptions, but used a fixed 10-word context instead.

Key Fields

(Additionally, there are various primary and foreign keys.)

Expression Indexing

Ordered Paths

Encode all vertical paths to the leaves, including the left-to-right position of each node
(Equivalently, encode all vertical paths from root to leaves, repeat recursively for each non-trivial subtree)

Expression Indexing

Ordered Paths

Encode all vertical paths to the leaves, including the left-to-right position of each node
(Equivalently, encode all vertical paths from root to leaves, repeat recursively for each non-trivial subtree)

Unordered Paths

Same, but without the position information

Expression Indexing

Ordered Paths

Encode all vertical paths to the leaves, including the left-to-right position of each node
(Equivalently, encode all vertical paths from root to leaves, repeat recursively for each non-trivial subtree)

Unordered Paths

Same, but without the position information

Sisters

Collection of sister nodes in the same subtree

Example Encoding

Expression

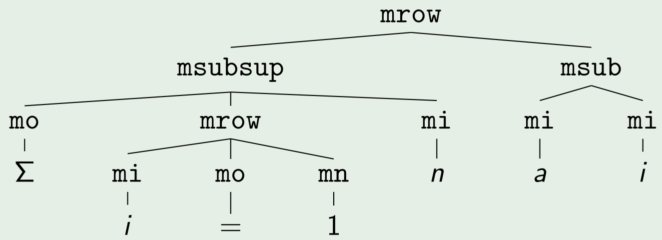
the polynomial $\sum_{i=1}^{n} a_i x^i$

Example Encoding

Expression

the polynomial $\sum_{i=1}^{n} a_i x^i$

MathML Tree

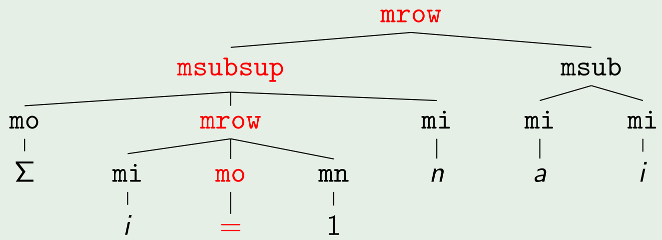


Example Encoding

Expression

the polynomial $\sum_{i=1}^{n} a_i x^i$

MathML Tree



Example opaths

1#2#2#mo#=

Example Encoding

Encoding

opaths: 1#msubsup 1#1#mo#Σ 1#2#1#mi#i 1#2#2#mo#= 1#2#3#mn#1 1#3#mi#n 2#msub 2#1#mi#a 2#2#mi#i
opaths: msubsup 1#mo#Σ 2#1#mi#i 2#2#mo#= 2#3#mn#1 3#mi#n
opaths: 1#mi#i 2#mo#= 3#mn#1
opaths: msub 1#mi#a 2#mi#i
upaths: #msubsup ##mo#Σ ###mi#i ###mo#= ###mn#1 ##mi#n #msub ##mi#a ##mi#i
upaths: msubsup #mo#Σ ##mi#i ##mo#= ##mn#1 #mi#n
upaths: #mi#i #mo#= #mn#1
upaths: msub #mi#a #mi#i
sisters: mi#i mo#= mn#1
sisters: mo#Σ mi#n
sisters: mi#a mi#i
sisters: msubsup msub
sisters: msubsup msub
description_en: the polynomial (indexed as: polynomi)

Querying

Querying

Encode the query MathML and description (whichever is provided) in the same way

Perform a disjunctive query on Lucene

Lucene scores matching terms, modified by tf/idf and length normalization

Querying

Querying

Encode the query MathML and description (whichever is provided) in the same way

Perform a disjunctive query on Lucene

Lucene scores matching terms, modified by tf/idf and length normalization

Document Queries

Document queries (with author, title or abstract terms) are done using a Lucene join query.

Querying

Querying

Encode the query MathML and description (whichever is provided) in the same way

Perform a disjunctive query on Lucene

Lucene scores matching terms, modified by tf/idf and length normalization

Document Queries

Document queries (with author, title or abstract terms) are done using a Lucene join query.

Jokers

Jokers implemented by simply leaving off the term from the query.

Alternately, due to the flexibility of the encoding, a single wrong leaf will only slightly penalize the score, not reject the result.

Results

	P-10 avg	**P-5 avg**	**MAP avg**	**Pre-cision**
Formula Search				
Relevant	0.229	0.219	0.162	0.065
Partially Relevant	0.500	0.476	0.379	0.220
Fulltext Search				
Relevant	0.293	0.320	0.297	0.103
Partially Relevant	0.660	0.680	0.534	0.309

Math Retrieval Conclusions

Conclusions

The current approach achieves satisfactory recall (for the first effort).

Precision is low, as there is no clear cutoff where the results stop being relevant.

Term-based search somewhat mitigates inconsistencies in representation.

Math Retrieval Conclusions

Conclusions

The current approach achieves satisfactory recall (for the first effort).
Precision is low, as there is no clear cutoff where the results stop being relevant.
Term-based search somewhat mitigates inconsistencies in representation.

Overall Conclusion

The approach, although simple, seems fruitful, and we intend to continue refining it.

Future Work

Future Work

Normalization of commonly interchangeable MathML elements.

Usage of Content MathML instead of Presentation MathML.

Giving more weight to operators and structure.

Implementation of common subexpression unification rules, which would additionally penalize the results where the instances of the same subexpression are replaced by different subexpressions.

Restriction of the number of disjunct terms, since their number adversely impacts search times.

Usage of actual *pq*-grams.

Usage of extracted descriptions.

Post-search reordering of top results using a more precise similarity measure.

Extraction of more advanced features for the math understanding subtask, such as information from dependency trees.

The End

Thank you for your attention