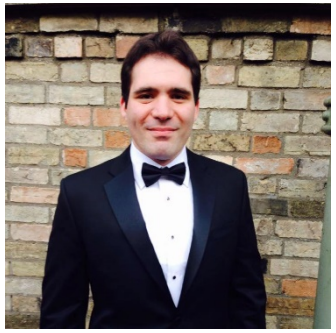
UNIVERSITY OF CAMBRIDGE

Variable Typing: Assigning Meaning to Variables in Mathematical Text

Marek Rei
marek.rei@cl.cam.ac.uk

Contributors



Yiannos Stathopoulos

Simon Baker

Marek Rei

Simone Teufel

Overview

- The task of variable typing
- The dataset for variable typing
- Intrinsic evaluation
- Extrinsic evaluation: mathematical IR

Introduction

- Texts from many major fields of study heavily rely on mathematics to communicate ideas and results.

- There's often an "interaction" of two contexts:
  - The textual context (flowing text)
  - Mathematical context (symbols and formulae).

- In this work, we introduce a new task focusing on a particular interaction between these two contexts:

  the assignment of meaning to variables by surrounding text.

Introduction

What is a “type” ?

- Multi-word phrases drawn from the mathematical technical terminology (Stathopoulos and Teufel, 2016)
- Types refer to
  - mathematical concepts (e.g., shape, number)
  - objects (e.g., matrix, set)
  - algebraic structures (e.g., group, ring)
  - physical concepts (e.g., energy, temperature).
- Typically noun phrases

The Variable Typing Task

Objective: Assign types to variables that appear in maths or scientific text.

For example:

An error of a single qubit can be expressed as a sum of operators taken from

the set $\rho = \{I, \sigma_x, \sigma_z, \sigma_x\sigma_z = i\sigma_y\}$, where I is the identity (corresponding to no error)

and σ_i are the Pauli spin operators

The Variable Typing Task

Objective: Assign types to variables that appear in maths or scientific text.

For example:

An error of a single qubit can be expressed as a **sum** of **operators** taken from

the **set** $\rho = \{I, \sigma_x, \sigma_z, \sigma_x\sigma_z = i\sigma_y\}$, where I is the **identity** (corresponding to no error)

and σ_i are the **Pauli spin operators**

The Variable Typing task

Objective: Assign types to variables that appear in maths or scientific text.

For example:

An error of a single qubit can be expressed as a **sum** of **operators** taken from the **set** $\rho = \{I, \sigma_x, \sigma_z, \sigma_x\sigma_z = i\sigma_y\}$, where I is the **identity** (corresponding to no error) and σ_x are the **Pauli spin operators**

The Variable Typing Task

Objective: Assign types to variables that appear in maths or scientific text.

For example:

An error of a single qubit can be expressed as a **sum** of **operators** taken from the **set** $\rho = \{I, \sigma_x, \sigma_z, \sigma_x\sigma_z = i\sigma_y\}$, where I is the **identity** (corresponding to no error) and σ_i are the **Pauli spin operators**

The Variable Typing Task

Why is this task useful ?

- Downstream tasks:
  - Mathematical information retrieval (MIR)
  - Topic modelling
  - Symbol disambiguation
  - Plagiarism detection
  - Open response mathematical questions (Lan et al., 2015)

- Can also be viewed as a relation extraction (RE) task.
  - Out of domain dataset for evaluating RE models and systems.

The Variable Typing Task

We make four key assumptions/constraints about this task:

1. Typing occurs at the sentential level and variables in a sentence can only be assigned a type phrase occurring in that sentence.

2. Variables and types in the sentence are known a priori.

3. Type relations in a sentence are independent of one another.

4. Type relations in one sentence are independent of those in other sentences – given a variable v in sentence s, type assignment for v is agnostic of other typings involving v from other sentences.

Detecting Types

- Extend seed dictionary by Stathopoulos and Teufel (2016)
  from 10,601 types to 1.2 million types.

- Basic idea:
  use seed dictionary to determine known “supertypes”, then return candidate phrases with known types in their suffix.

- Example
  - known types: {algebra, lie algebra, tensor, riemannian manifold}
  - expanded types: {coalgebra, cotensor, complete riemannian manifold, submanifold, isotropic submanifold, order cotensor}

The Variable Typing Data Set

We constructed a corpus for variable typing:

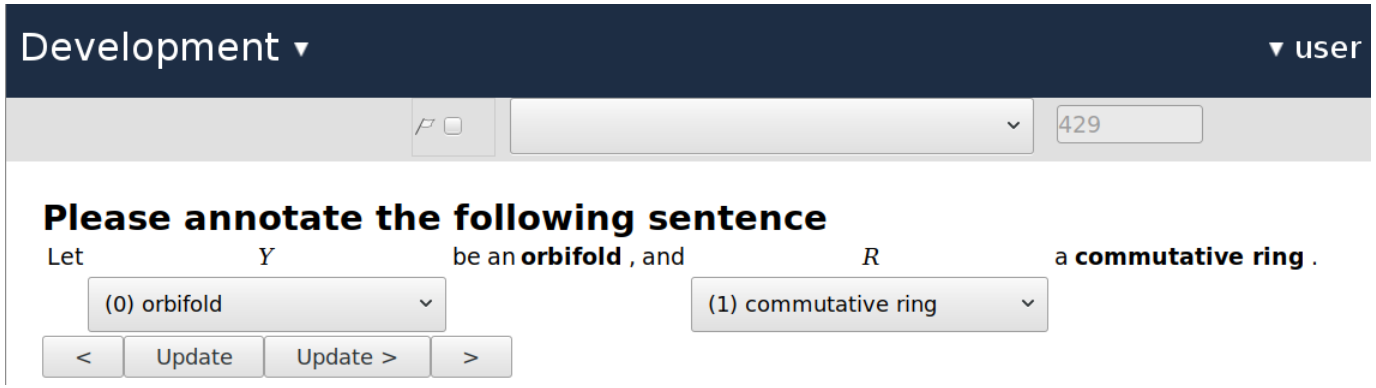
- Using text from scientific articles in the MREC corpus (Líška et al., 2011)
- Identify variables in text using Symbol Layout Trees from MathML
- Select sentences for annotation, using stratified sampling based on the location of the sentence (intro, theorems, etc) and the number of variables.

	Train	Dev	Test	Total
Sentences	5,273	841	1,689	**7,803**
Positive edges	1,995	457	1,049	**3,501**
Negative edges	15,164	4,386	10,473	**30,023**
Total edges	17,159	4,843	11,522	**33,524**

The Variable Typing Data Set

Human annotation and agreement:

- 2 annotators for agreement experiment
- 108 sentences/182 relations overlap



Relation Labels

Varying labels

- One label per type instance

Fixed labels

- Type Unknown
- Type Present but Undetected
- Parameterisation
- Index
- Number
- Formula is not a variable

The Variable Typing Data Set

Human Annotation and Agreement

- 2 annotators for agreement experiment
- 108 sentences/182 relations overlap

Results:

- Annotators agree that a variable can be typed by its context?
  - Cohen's Kappa 0.80 (substantial)
- Given a variable is typable from its context, do the annotators agree on the type?
  - Accuracy 90.9%
- Annotators agree that a variable is not typable
  - Cohen's Kappa 0.61 (moderate)

Intrinsic Evaluation

Objective:
Evaluate machine learning algorithms on the task of variable typing

Models / baselines:

- Nearest type (naïve baseline, assign the nearest type)
- SVM model by Kristiano et al. (2012), using hand-written rules as features
- SVM+, adding features specific to our task
- CNN model, classifying each possible edge
- BiLSTM, formulated as a sequence labeling task

Intrinsic Evaluation - Results

Model	Precision	Recall	F1-Score
Nearest type	30.30	82.94	44.39
SVM	55.39	76.36	64.21
SVM +	71.11	72.74	71.91
CNN	80.11	70.26	74.86
BiLSTM	**83.11**	**74.77**	**78.98**

Extrinsic Evaluation – MIR

Mathematical Information Retrieval (MIR):

- Retrieval of mathematical information needs, expressed as text and formulae.
- Using the Cambridge University MathIR test collection (CUMTC)
- Queries and relevant documents are rich in mathematical types

Example query:

Let P be a parabolic subgroup of $GL(n)$ with Levi decomposition $P = MN$, where N is the unipotent radical.

Let π be an irreducible representation of $M(\mathbf{Z}_p)$ inflated to $P(\mathbf{Z}_p)$,

how does $Ind_{P(\mathbf{Z}_p)}^{GL_n(\mathbf{Z}_p)} \pi$ decompose?

It would be sufficient for me to know the result in the simplest case, where P is a Borel subgroup

Example relevant document:

BRANCHING RULES FOR UNRAMIFIED PRINCIPAL SERIES REPRESENTATIONS OF GL(3) OVER A p-ADIC FIELD

PETER S. CAMPBELL AND MONICA NEVINS

ABSTRACT. On restriction to the maximal compact subgroup $GL(3, \mathfrak{R})$, an unramified principal series representation of the p-adic group $GL(3, F)$ decomposes into a direct sum of finite-dimensional irreducibles each appearing with finite multiplicity. We describe a coarser decomposition into components which, although reducible in general, capture the equivalences between the irreducible constituents.

1. INTRODUCTION

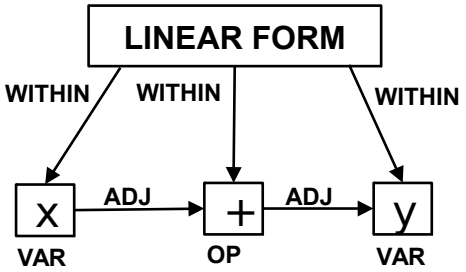
The aim of this paper is to investigate the relationship between the representation theory of a p-adic group G and its maximal compact subgroups K. Given an admissible representation of G, its restriction to K decomposes as a direct sum of smooth irreducible representations of K each with finite mul-

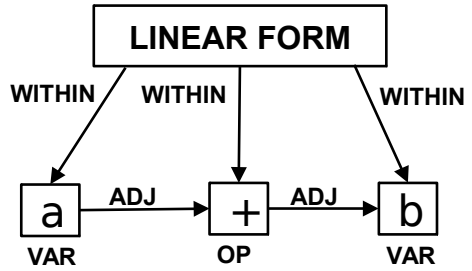
Extrinsic Evaluation – MIR

Intuition: Enriching formulas with types from the text can be beneficial for MIR.

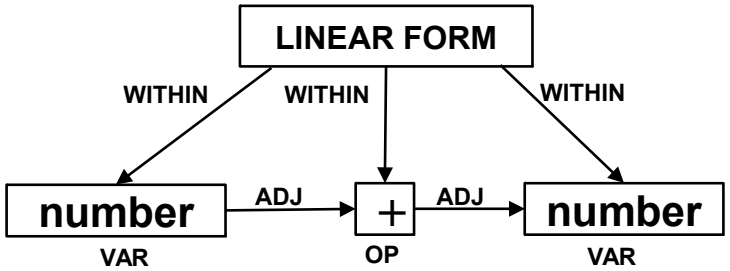
- Formulas are represented using Symbol Layout Trees (SLTs)

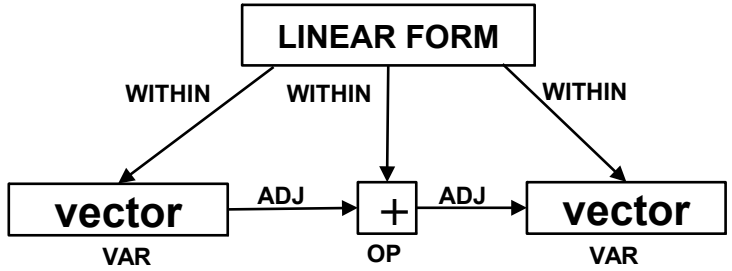
- For example, formulas x + y and a + b are represented is a similar way:





- However, if x,y may have type 'number', while a,b have type 'vector':





Extrinsic Evaluation - Results

Using math-specific IR system Tangent (Pattaniyil et. al, 2014)
Measuring mean average precision (MAP)

Standard IR baselines

Tangent Index	
VSM	BM25
.076	.079

Math-specific IR models

	Tangent Index	
	Untyped formulas	Typed formulas
No types in text	.052	.046
Types in text	.083	**.139**

Bold: difference statistically significant with all models at $p < 0.05$

Conclusions

1. We introduced the new task of variable typing
2. We have produced a new data set of 33,524 data points
3. Trained ML models for variable typing, with a BiLSTM sequence labeler performing best
4. Our extrinsic evaluation demonstrates that the data set is useful for downstream tasks
5. We make our variable typing data set available through the open data commons license

For more information: http://www.cl.cam.ac.uk/~yas23/

Questions ?

Thank you!