Building Dataset for Grounding of Formulae

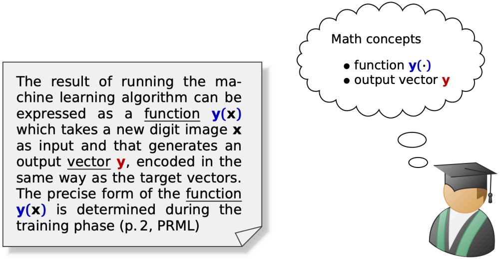
— Annotating Coreference Relations Among Math Identifiers —

Takuto Asakura, Yusuke Miyao, Akiko Aizawa

LREC 2022 (prerecord)

Grounding of Formulae [Asakura+ 2020]

1. Finding groups of tokens which refer to math concepts
   E.g. x, α, cos, $\sum$, $=$, $\times$, etc.
2. Associating a corresponding math concept to each group



Our contribution: Built a dataset for automating the grounding

- Manually annotated **12,352 math identifiers** in 15 papers
- Revealed **scope switch of identifiers is frequent and complex**

Grounding of Formulae [Asakura+ 2020]

≈ *Description Alignment* + Coreference Resolution

- A task to associate description for each math identifier occurrence
- There are some existing work [Aizawa+ 2013, Alexeeva+ 2020, etc.]

the generalization loss

the output of predicator $\hat{t}$ for an input x

a random variable for a test input for a regression problem

a true joint distribution in general, without any specific definition

$$L_p(\hat{t}) = \mathrm{E}_{(x,t)\sim p(x,t)}\left[\ell\left(t, \hat{t}(x)\right)\right]$$

a given loss function

a random variable for a test output for a regression problem

the average with the condition

the true joint distribution itself

Grounding of Formulae [Asakura+ 2020]

≈ Description Alignment + *Coreference Resolution*

Coreference in Natural Languages

Bob told Alice that

Coreference

he wants to study NLP.

Coreference in Formulae

The result of running the machine learning algorithm can be expressed as a function $\mathbf{y}(\mathbf{x})$ which takes a new digit image $\mathbf{x}$ as input and that generates an output vector $\mathbf{y}$, encoded in the same way as the target vectors. The precise form of the function $\mathbf{y}(\mathbf{x})$ is determined during the training phase (PRML, p. 2)

Difficulty and Necessity of Formulae Grounding

- Various ambiguities similar to natural languages [Kohlhase+, 2014]
  - A symbol (token) can be used in several meanings
  - Syntactic ambiguity E.g. $f(a+b)$
- Formulae cannot be understood without reading surrounding texts
- Common sense and domain knowledge may be required
  E.g. π is Archimedes' constant

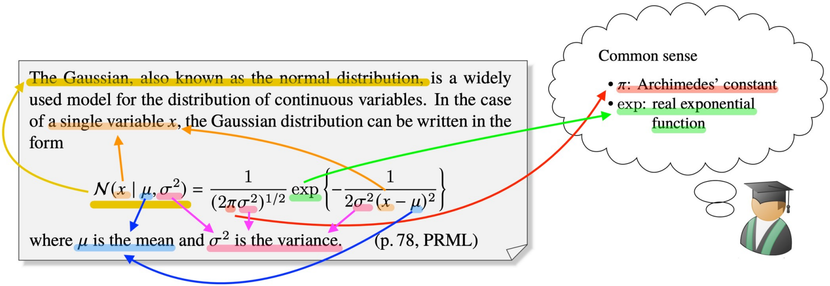
Usage of character y in the first chapter of PRML (except exercises)

Text fragment from PRML Chap. 1	Meaning of $\mathbf{y}$
... can be expressed as a function $\underline{\mathbf{y}}(\mathbf{x})$ ...	a function which takes an image as input
... an output vector $\underline{\mathbf{y}}$, encoded in ...	an output vector of function $\mathbf{y}(\mathbf{x})$
... two vectors of random variables $\mathbf{x}$ and $\underline{\mathbf{y}}$ ...	a vector of random variables
Suppose we have a joint distribution $p(\mathbf{x}, \underline{\mathbf{y}})$ ...	a part of pairs of values, corresponding to $\mathbf{x}$

Source of Grounding (SoG)

Bases of grounding of formulae inside or outside documents:

inner Surrounding texts, formulae E.g. apposition noun, $\stackrel{\text{def}}{=}$
outer Common sense, domain knowledge E.g. Wikidata



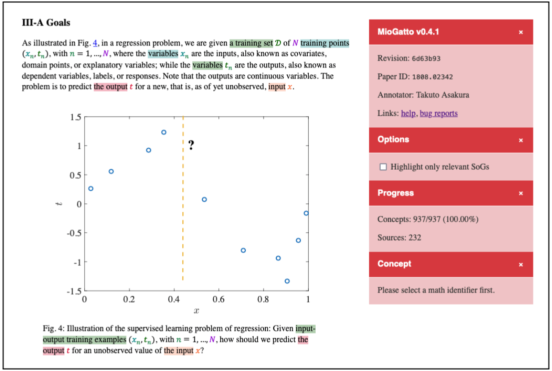
Things annotated — Information that will be needed for automation

- **Math concepts** are **the ground truth** of the grounding
- **Sources of grounding** will be extracted first for automating

MioGatto — The Annotation Tool [Asakura+ 2021]

Math Identifier-Oriented Grounding Annotation Tool

- Special annotation tool for building our grounding dataset
- Available as an **open source software** (MIT license)



https://github.com/wtsnjp/MioGatto

Annotation Method

Annotators

We recruited **10 student annotators** (paid)

- in various fields:
  NLP × 4, Logics × 2, Mathematics × 1, Physics × 1, Astronomy × 1
- in various grades:
  high school × 1, undergrad × 1, Master × 5, Doctoral × 3

Method

- Annotation targets are **math identifiers** E.g. x, θ, sin
- The target papers are basically selected by annotators
- **Annotation guideline** is provided for the annotators

https://github.com/wtsnjp/MioGatto/wiki/Annotator's-Guide

Annotation Results — Dataset Overview

Dataset for formulae grounding

No.	Domain	#words	#types	#occr	#concepts	**Avg. #candidates**	#sources
1	ML	10976	40	937	104	**6.4**	232
2	NLP	4267	42	266	73	2.6	30
3	NLP	3563	38	433	79	2.5	34
4	Logics	3567	46	1648	64	1.9	30
5	Algebra	13154	141	4629	424	5.2	180
6	NLP	2881	25	162	30	2.7	12
7	NLP	5543	31	203	47	2.6	36
8	NLP	4613	23	217	27	**1.1**	28
9	NLP	6255	34	510	74	2.7	27
10	NLP	5415	73	1175	167	3.3	60
11	NLP	4451	33	237	61	2.9	34
12	NLP	4261	31	186	39	1.7	25
13	NLP	2257	23	124	27	1.2	18
14	Astronomy	10032	59	1064	129	4.2	97
15	Astronomy	4863	41	561	73	2.3	95
Sum	—	86098	680	**12352**	1418	—	**938**

https://sigmathling.kwarc.info/resources/grounding-dataset/

Dataset Analysis (1) Inter-annotator agreements

Inter-annotator agreements (to Annotator A)

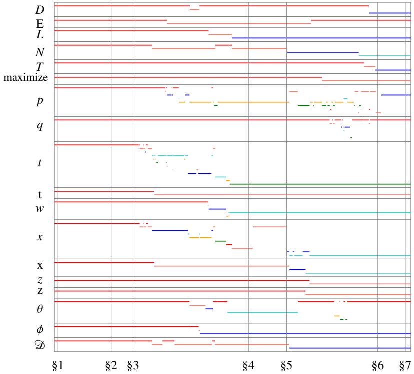
Annotator	A	B	C	D	E
Agreement (%)	—	**96.5**	87.4	92.1	**84.2**
Cohen's κ^*	—	**0.94**	0.80	0.87	**0.75**
#SoGs	232	—	—	249	257
Overlap (%)	—	—	—	**80.3**	**93.4**

* Weighted average according to the #occr

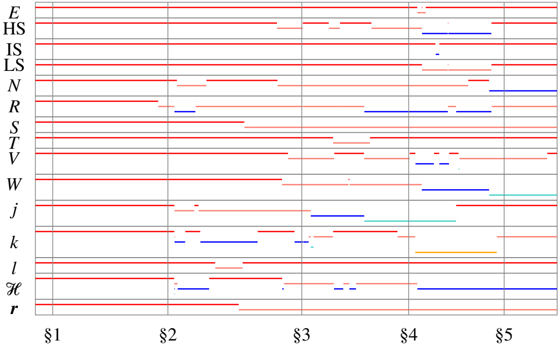
- Five people independently annotated Paper 1
  - Mah concepts are annotated by all
  - Sources are annotated by Annotator A, D, E
- Both agreements and Cohen's κ for math concepts are **high**
- Text spans that are recognized as SoGs are **hevily overlap**

Dataset Analysis (2) Scope Switches

Paper 1



Paper 15



Scope switches — changes of math identifier meanings

- 89.5% of them occur within a single section
- The scopes of identifier can back and forth

Dataset Analysis (3) Source of Grounding

Examples of grounding sources

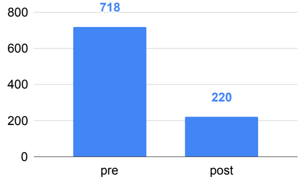
> In the case of **a single variable** x, the Gaussian distribution can be written. . . (p. 78, PRML)

Analyses on annotated 938 SoGs

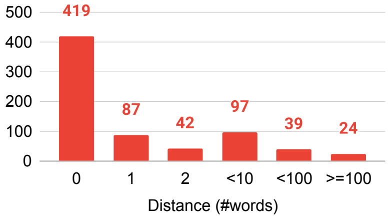
- 76.5% of them are **pre** SoG
- Distance between identifier and SoG is 14.7 words in average
  cf. **Median is 0–4**

Typical SoGs are **apposition nouns**

Position of SoG



Identifier — SoG distance



Future Work

Reducing annotation costs

- Difficult to annotate a paper by multiple annotators
  → we could not get inter-annotator agreements for all papers
- Still not enough data to compare among different domains
  - **Too many math formulae** in papers about Mathmatics and Physics
    → We need some automation. **Create only dictionaries first**
  - **Notations are especially trickey** in papers for math logics
    → **Disambiguation for numbers and operators are needed**

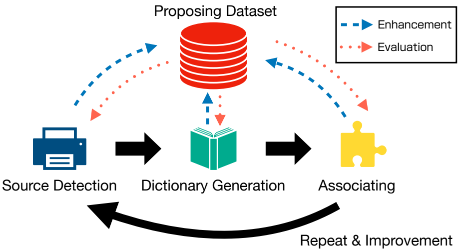
Further unanswerd research questions

- Are there differences between annotation by authors and readers?
- Can people who are not specialized for the domain also perform the annotation?

The Strategy for the Grounding Automation

3-step of Automation

1. **Detecting/Retrieving** inner-document sources of grounding
   → Pattern matching + POS tagging
2. **'Dictionary' generation** by clustering the sources
   → Short text clustering [Jiaming+, 2017] may be applicable
3. **Associating each occurrence** with the entry in the 'dictionary'
   → Pattern matching + POS tagging + text classification



References

- Akiko Aizawa, et al. "NTCIR-10 Math Pilot Task Overview." In *Proceedings of NTCIR-10* (2013).
- Maria Alexeeva, et al. "MathAlign: Linking Formula Identifiers to their Contextual Natural Language Descriptions". *Proceedings of LREC 2020*.
- Takuto Asakura, et al. "Towards Grounding of Formulae.". In *Proceedings of SDP 2020*.
- Takuto Asakura, et al. "MioGatto: A Math Identifier-oriented Grounding Annotation Tool." In *13th MathUI Workshop at 14th Conference on Intelligent Computer Mathematics (MathUI 2021)*.
- Christopher M Bishop. *Pattern Recognition and Machine Learning* (2006).
- Xu, Jiaming, et al. "Self-taught convolutional neural networks for short text clustering." *Neural Networks 88* (2017).
- Michael Kohlhase and Mihnea Iancu. "Co-representing structure and meaning of mathematical documents" (2014).