Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4851–4858
Marseille, 20-25 June 2022

Building Dataset for Grounding of Formulae — Annotating Coreference Relations Among Math Identifiers

Takuto Asakura[1], Yusuke Miyao[1], Akiko Aizawa[1,2]
[1]Dept. of Computer Science, The University of Tokyo, [2]National Institute of Informatics, Tokyo, Japan
{takuto,yusuke}@is.s.u-tokyo.ac.jp, aizawa@nii.ac.jp

Abstract

Grounding the meaning of each symbol in math formulae is important for automated understanding of scientific documents. Generally speaking, the meanings of math symbols are not necessarily constant, and the same symbol is used in multiple meanings. Therefore, coreference relations between symbols need to be identified for grounding, and the task has aspects of both description alignment and coreference analysis. In this study, we annotated 15 papers selected from arXiv.org with the grounding information. In total, 12,352 occurrences of math identifiers in these papers were annotated, and all coreference relations between them were made explicit in each paper. The constructed dataset shows that regardless of the ambiguity of symbols in math formulae, coreference relations can be labeled with a high inter-annotator agreement. The constructed dataset enables us to achieve automation of formula grounding, and in turn, make deeper use of the knowledge in scientific documents using techniques such as math information extraction. The built grounding dataset is available at `https://sigmathling.kwarc.info/resources/grounding-dataset/`.

Keywords: Math Linguistics, Math Information Retrieval (MathIR), Coreference Relations, Annotation Tool

1. Introduction

Understanding math formulae is as important as understanding natural language texts to analyze documents in science, technology, engineering, and mathematics. Analyzing math formulae is unavoidable to fully exploit the knowledge contained in scientific documents by using applied technology in computer science such as information retrieval, computer algebra systems, and theorem proving. In order to understand a math formula in documents, it is necessary to clarify the meaning of each formula token, that is, a character or symbol that appears in the math formula. This part of formula analysis is formalized as a task of formula grounding (Asakura et al., 2020). The grounding task has two characteristics: one is the description alignment task, which assigns a context-specific description to each formula token (Figure 1), and the other is the coreference resolution task, which discriminates tokens that are used with exactly the same meaning from those that are not.

In order to automate the process of formula grounding, the authors first worked on constructing a corpus with ground truth annotations manually for observation, analysis, learning, and evaluation. As annotation of coreference information is generally costly (Oberle, 2018), the authors developed a special annotation tool, MioGatto[1], to streamline the data construction process (Asakura et al., 2021). The authors then used MioGatto to annotate a total of 15 scientific papers with 11 student annotators for all occurrences of math identifiers in the papers.

In this paper, we introduce the procedure of constructing a dataset of formula grounding and report an overview of the constructed annotated corpus.

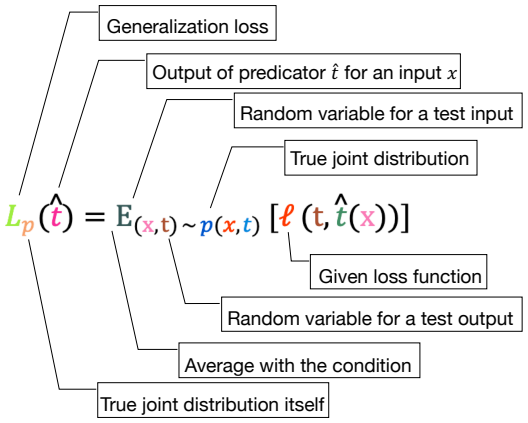

Figure 1: Description alignment

2. Related Work

The arXMLiv dataset (Ginev, 2020) is a large corpus of more than 1.5 million scientific papers on the preprint server arXiv.org[2], converted into XHTML documents using LaTeXML (Miller, 2018) for easy handling by computer programs for various research purposes. In the documents, math formulae are mechanically converted to presentation MathML (Ausbrooks et al., 2014) by LaTeXML, but essentially the same information as LaTeX, about what the formula looks like, is encoded, without any additional information. The arXMLiv dataset is widely used as a valuable linguistic resource for documents containing mathematical expressions, and the input format of MioGatto follows the XTHML specification of the dataset (Ginev et al., 2011).

Several annotated corpora of scientific papers have been proposed, in which each token of a formula is given a

[1]`https://github.com/wtsnjp/MioGatto`

[2]`https://arxiv.org`

description. In NTCIR-10, a subtask of Math Understanding was proposed to extract definitions of tokens in natural language text as part of the Math Pilot Shared Task. A dataset of manually annotated math formulae in XHTML documents included in the arXMLiv dataset was provided for development and evaluation for the task (Aizawa et al., 2013). The MathAlign task was also formulated as a similar task that assigns an explanation to each math identifier in formulae, and a dataset of 584 math identifiers from 116 papers in the arXiv.org collection is also constructed (Alexeeva et al., 2020).

In real-world scientific documents of certain length, symbols and letters in math formulae are often used in multiple meanings within a single document (Asakura et al., 2020). For example, in Chapter 1 of Pattern Recognition and Machine Learning (PRML) (Bishop, 2006), a textbook in the field of machine learning, the bolded **y** is used in at least four different meanings in the same chapter (Table 1). Therefore, to understand math formulae in a document, it is necessary to resolve the coreference relations among these tokens in the same document. However, there is no known dataset that explicitly labels coreference relations between tokens of math formulae. In this study, we selected 15 scientific papers, mainly those with sufficient amount of math formulae, and annotated all 12,352 occurrences of math identifiers in the papers so that the coreference relations within each paper are explicit.

3. Purpose and Method

Datasets are fundamental to the construction and evaluation of methods for automated formula grounding. Large amounts of training data are generally required to build a statistical model for such automation. Although we plan to use a rule-based method to increase the amount of data initially, we still need some amount of manually annotated ground-truth data, as we have to observe the usage of formula tokens in real documents to study the rules. We manually annotated the following two types of information for actual scientific papers as a first step to automate formula grounding (Figure 2).

Math concepts that formula tokens refer to. In terms of actual annotation data, additional attributes such as mathematical type, arity, and constraints can be added to the simple descriptions.

Sources of grounding, text spans that can be used as bases for human to ground formula tokens. Mathematically, a source of grounding is a definition or declaration. For example, the first f in Figure 2 is grounded to a real-valued function, and the source of grounding is the preceding "a real-valued function."

Instead of directly annotating each occurrence of a math token with a description, we annotated each token with a concept ID defined in the *math concept dictionary*, a list of math concepts created by an annotator. This enables us to construct a dataset with explicit coreference relations between formula token occurrences: those associated with the same concept ID have a coreference relation, while those associated with different IDs have no coreference relation. To achieve such annotation, we used MioGatto, a special annotation tool developed by the authors (Figure 3). MioGatto is designed to easily (1) create a math concept dictionary, (2) assign a math concept ID to each occurrence of a formula token, and (3) annotate the span position of the grounding source using only intuitive GUI operations (Asakura et al., 2021).

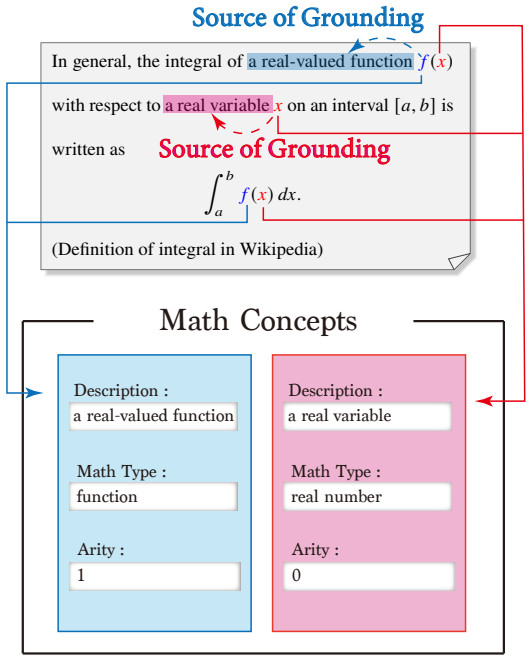

Figure 2: Two types of information we annotated: math concepts and sources of grounding. The example sentence is taken from Wikipedia[4].

In order to construct the dataset, we selected and annotated papers that contain more than a certain amount of math formulae from papers available on arXiv.org with LaTeX document sources. Since reading and annotating such specialized scientific papers with math formulae requires specialized knowledge in appropriate fields, we recruited collaborators from among students (mainly graduate and undergraduate students) with specialized knowledge in a variety of fields. Majority of the annotators we collected were specialized for natural language processing, with others majoring in mathematical logic, algebra, physics, and astronomy. We asked the participating annotators to select papers from the arXiv.org collection that matched their background knowledge and interests, and preprocessed the selected papers for the annotation using MioGatto. This preprocessing includes converting the original LaTeX doc-

[4] https://en.wikipedia.org/wiki/Integral

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 $\mathbf{x}$

Table 1: Meanings of $\mathbf{y}$ in Chapter 1 of PRML (Bishop, 2006).

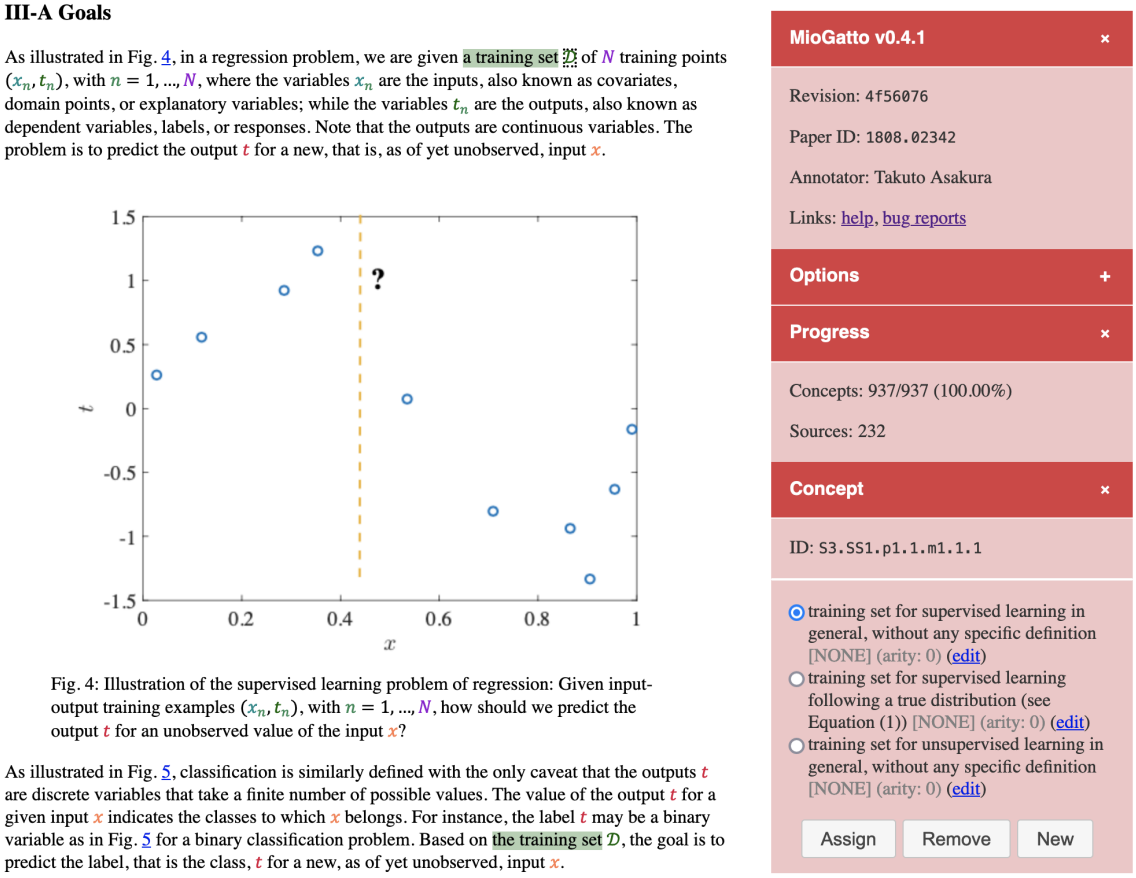

Figure 3: Screenshot of MioGatto when annotating an arXiv paper in the field of machine learning (Simeone, 2018). The left side of the screen contains the text of the article to be annotated, and the right side contains the information provided by MioGatto and the buttons necessary for the annotation operation.

uments by the authors into XHTML with LaTeXML, and correcting erroneous markup in math formulae by the original authors of the target papers. Each annotator was provided with a guideline[5] on how to use MioGatto, as well as XHTML data and annotation data templates for the actual annotation. The annotators performed the annotation work following the guideline. After that, the data obtained from the annotation was checked by the authors and analyzed.

The target of this annotation is the all occurrences of math identifiers for all math formulae used in the selected papers. A math identifier is a kind of formula token, which is a single letter (e.g., x and θ) or a short name (e.g., sin) representing a variable, function, or constant. Technically, math tokens that appear as `<mi>` tags in presentation MathML are annotated. There are other formula tokens such as operators (e.g., +) and numbers in math formula, but we focus on math identifiers to avoid too many annotation targets. We did not limit the number of grounding sources because there may be multiple sources associated with a concept or no sources associated with a concept in a document.

4. Analysis for the Dataset

We completed the manual annotation of all math identifiers in 15 scientific papers in the fields of natural language processing, mathematical logic, algebra, and astronomy (Table 2 and Table 3). In total, there were 12,352 occurrences of math identifiers in the entire

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

No.	Author	Title	arXiv ID	arXiv category
1	Osvaldo Simeone	A Very Brief Introduction to Machine Learning With Applications to Communication Systems	1808.02342	cs.IT
2	Tsung-Hsien Wen et al.	Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems	1508.01745	cs.CL
3	Qian Chen et al.	Enhanced LSTM for Natural Language Inference	1609.06038	cs.CL
4	Joseph Singleton	A Logic of Expertise	2107.10832	cs.LO
5	Edward Frenkel	Recent Advances in the Langlands Program	math0303074	math.AG
6	Laura Aina et al.	Putting words in context: LSTM language models and lexical ambiguity	1906.05149	cs.CL
7	Jian Guan et al.	A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation	2001.05139	cs.CL
8	Richard Antonello et al.	Selecting Informative Contexts Improves Language Model Finetuning	2005.00175	cs.CL
9	Jinhua Zhu et al.	Incorporating BERT into Neural Machine Translation	2002.06823	cs.CL
10	Xuan-Phi Nguyen et al.	Tree-structured Attention with Hierarchical Accumulation	2002.08046	cs.CL
11	Jiangang Bai et al.	Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees	2103.04350	cs.CL
12	Zenan Xu et al.	Syntax-Enhanced Pre-trained Model	2012.14116	cs.CL
13	Yangyifan Xu et al.	Bilingual Mutual Information Based Adaptive Training for Neural Machine Translation	2105.12523	cs.CL
14	Daisuke Taniguchi et al.	Effective temperatures of red supergiants estimated from line-depth ratios of iron lines in the YJ bands, 0.97–1.32 micron	2012.07856	astro-ph.SR
15	Daisuke Taniguchi et al.	Pressure-induced two-step spin crossover in double-layered elastic model	1708.02771	cond-mat.mtrl-sci

Table 2: The reference information of papers in our annotated dataset.

dataset, all of which were assigned math concepts. Altogether, 938 text spans were also collected, which are sources of grounding. By dividing the number of math identifier-types in each paper by the number of math concepts in the corresponding dictionary, we can calculate the average number of meanings used for each math identifier-type, which is 2.09 for the whole dataset.

Given that the number of occurrences of each math identifier-type is different, the weighted average of the number of occurrences is the "Avg. #candidates" in Table 3. This corresponds to the average number of choices that the annotator sees when assigning a math concept to each occurrence during the actual annotation. Therefore, the higher the value of "Avg. #candidates", the higher the degree of math identifier ambiguity, and the higher the difficulty of the annotation. Since math identifiers are often single letters of the alphabet rather than descriptive names, there is a limit to the variety of math identifier-types that can be used, even taking into account differences in variants such as roman and calligraphy typefaces. For this reason, the longer the document is, i.e., the more words there are, the higher the average number of candidates and the stronger the ambiguity is (Figure 4). This shows the importance of annotating not only short documents or parts of long documents, but also entire long documents with large ambiguities, as we did here.

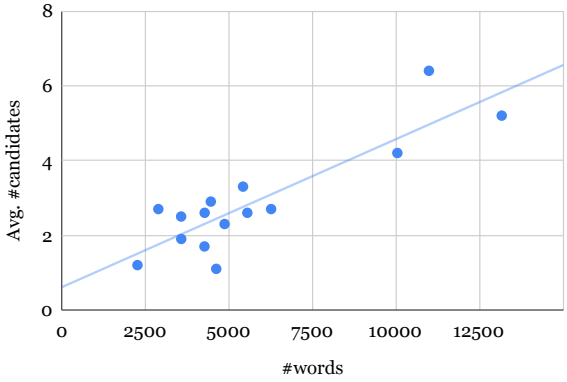

Figure 4: Relationship between number of words in the papers and the "Avg. #candidates". The correlation between the two is strong positive, with a correlation coefficient of $r = 0.87$.

4.1. Inter-annotator Agreements

Since the target of the annotation in this study is a highly specialized scientific paper, it is not easy to secure multiple annotators for the same paper. However, in order to confirm the accuracy and reproducibility of the annotation, a total of five annotators annotated Paper 1 independently of each other, and the inter-annotator agreement rate was calculated (Table 4). Annotator A was responsible for creating the math concept dictio-

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

Table 3: Annotation results. Herein, the leftmost column "No." is the paper ID for convenience of explanation, "#words" is the number of words in the text of the paper, "#types" is the number of used math identifier-types, and "#occurrences" is the number of math identifier occurrences. The next column "#concepts" is the number of math concepts in the concept dictionary. "Avg. #candidates" is the weighted average of the number of dictionary entries according to the number of identifier occurrences, and "#sources" is the number of grounding sources.

Annotator	A	B	C	D	E
Create concept dict.	✓				
Assign concepts	✓	✓	✓	✓	✓
Label sources	✓			✓	✓
Agreement rate (%)	—	96.5	87.4	92.1	84.2
Cohen's κ[6]	—	0.94	0.80	0.87	0.75
Number of sources	232	—	—	249	257
Overlap rate (%)	—	—	—	80.3	93.4

Table 4: Annotator roles and inter-annotator agreement rates. The top three rows show the role of each annotator, the middle two rows show the agreement rate of math concepts, and the bottom two rows show the information of grounding sources. The agreement rate and overlap rate were calculated between annotator A and each annotator in the others.

nary, while the other annotators used the dictionary to assign concepts and annotate the sources of grounding. Although the accuracy of the work varied slightly depending on the annotator, the agreement rate of math concepts and Cohen's κ (Cohen, 1960) were calculated to be high enough. Grounding sources also overlapped with high frequency (80.3–93.4%), meaning that text spans that humans consider as grounding sources were found to match well.

4.2. Math Concept Dictionaries

In this dataset construction, a dictionary of math concepts was created by an annotator for each of the 15 scientific papers. As shown in Table 3, a total of 1,418 math concepts for 680 math identifier-types were registered in 15 dictionaries. As a concrete example, Table 5 shows a portion of the dictionary created for Paper 1. Each concept has a short description of 6.7 words on average across all dictionaries and some additional attributes: the first ones are affixes, which contain information about the notation, such as whether they are accompanied by superscripts or not, and whether they have parentheses to represent the function's arguments. On average, the number of affixes registered for each concept was 0.8. The second one is arity, which is the information about how many arguments a concept semantically takes when it is a function. The dictionary we have constructed contains function concepts with arity 0 to 4.

For each math identifier-type, up to 14 math concepts are registered in the math concept dictionaries (Figure 5). In Table 6, we list the top 5 math identifier-types with the highest average number of math concepts.

4.3. Scope Switches

If the math concept assigned to an occurrence of a math identifier is distinct from the concept assigned to a previous occurrence of the same identifier-type, we say that there is a *scope switch* between the two occurrences of the math identifier. In order to perform formula grounding, we need to identify all scope switching locations in a single document, which is the most challenging part of the automation. The dataset we constructed contained a total of 2,378 scope switches throughout 15 papers. Of these, 2,129 (89.5%) occurred within a single section, indicating that there is ambiguity in the meanings

[6]Weighted average according to the number of occurrence of each math identifier-type.

Identifier	Description	Affixes	Arity
D	the number of dimensions for the vector x	(NONE)	0
	the number of dimension for the fixed features $\phi(x)$	prime	0
	f-divergence	subscript, open parenthesis, ...	2
	⋮		
t (*italic*)	an output of a regression or classification problem in general	(NONE)	
	an output of a regression problem, generated by $p(x, t)$	(NONE)	0
	n-th output in the training set $\mathcal{D}$	subscript	0
	a predicator which takes an input x and return a predicated value	over, parentheses	0
	⋮		
t (roman)	a random variable for a test output for regression problem	(NONE)	0
	⋮		

Table 5: Excerpt from the math concept dictionary for Paper 1.

Identifier-type	Example meaning	Avg. #concepts	Used in
$\boldsymbol{N}$	hidden representation	9.0	Paper 10
$\mathbf{W}$	parameters	8.0	Paper 2, 3
$\mathbf{U}$	parameters	8.0	Paper 3
$\mathbf{v}$	fixed-length vector	7.0	Paper 3
Bun	moduli space	7.0	Paper 5

Table 6: Math identifier-types that have many concepts in the math concept dictionaries.

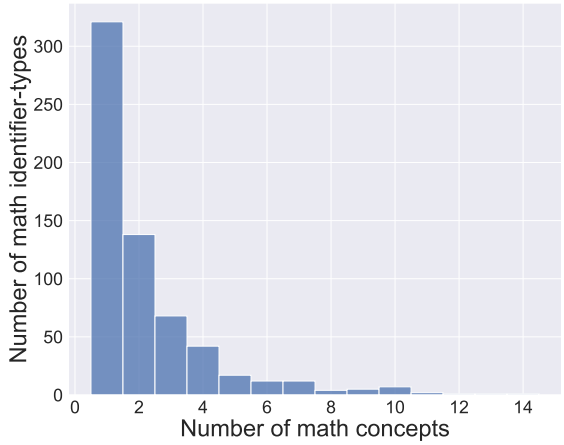

Figure 5: Number of math concepts for each entry (math identifier-type) in the math concept dictionaries.

of math identifiers even when we focus on the narrow scope of a single section in a document. It should be noted that the scope of math identifiers is complex: it often switches in units finer than sections and sometimes switches back to the original scope once it has switched to another scope (Figure 6).

4.4. Sources of Grounding

We also analyzed the positional relation between the textual span of the grounding sources and the occurrence of the math identifiers associated with them (Table 7). Out of the 938 grounding sources annotated, 718 (76.5%) were found before the corresponding math identifier occurrence. In terms of the number of words in between, the average distance between each grounding source and its associated math identifier occurrence that is closest was 14.7 words. However, distances between sources and math identifier occurrences vary widely, with a median distance ranging from 0 to 4 words across all papers. In a nutshell, typical sources of grounding are within a few words before matching the occurrences of math identifiers.

No.	Position		Distance (words)	
	Pre	Post	Mean	Median
1	217	15	0.3	0
2	28	2	1.8	0
3	19	15	19.9	2
4	18	12	4.1	1
5	105	75	1.2	0
6	9	3	35.0	1
7	31	5	20.5	4
8	19	9	8.9	0
9	23	4	2.1	3
10	57	3	3.6	0
11	29	5	9.4	4
12	20	5	17.2	3
13	16	2	0.3	0
14	64	33	75.9	3
15	63	32	30.4	2
Total	718	220	—	—

Table 7: Statistics of grounding sources in the dataset.

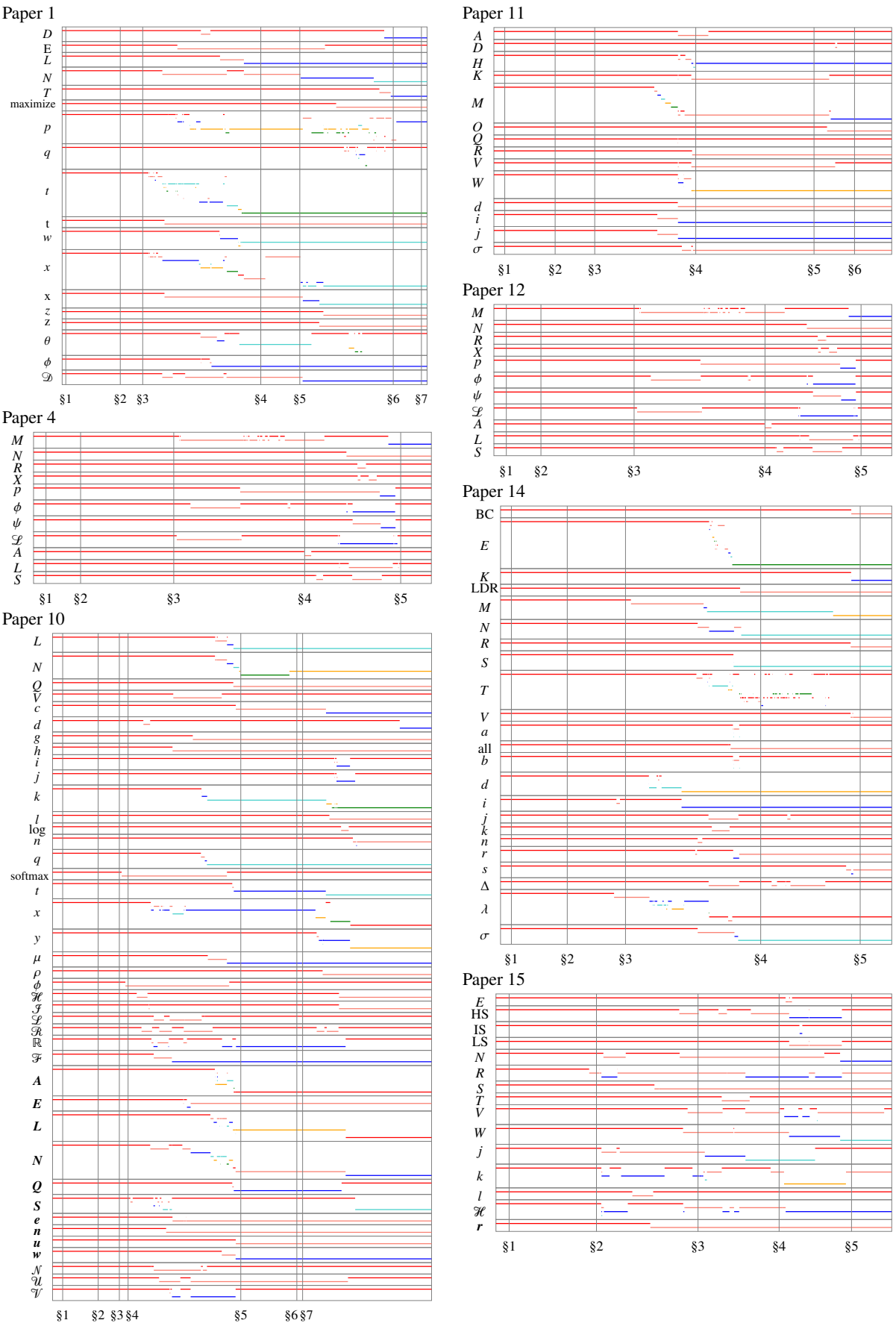

Figure 6: Scopes of math identifiers in the selected papers. The horizontal axis indicates the position within each paper, and the vertical axis indicates the math concept each scope corresponds to. Where the colored horizontal lines that represent the scopes are interrupted, it means that there are scope switches. To clearly indicate the position of scope switches, horizontal lines are drawn as such any position in the document belongs to a scope.

5. Conclusions and Future Work

In this study, we constructed a dataset of 15 scientific papers in various domains which were manually annotated with grounding information. Each occurrence of a math identifier in the dataset is labeled with a description and some additional information, and the coreference relations between math identifiers within each paper are made explicit. We also showed that such a dataset can be constructed by an annotator that is not necessarily specialized in constructing linguistic resources.

In the future, we will make up only a math concept dictionary by hand, and automatically assign the appropriate entry from the dictionary to each occurrence of a math identifier in the paper. In this way, the proposed dataset can be effectively extended quantitatively, and further, we accomplish the whole automation of the formula grounding process.

6. Acknowledgements

This work has been supported by JST, ACT-X Grant Number JPMJAX2002, Japan. We are grateful to Mr. Taiga Ishii for his bug reports and feedback on the tool. We would like to thank Prof. Michael Kohlhase, Mr. André Greiner-Petter, and Mr. Jan Frederik Schaefer for fruitful discussions.

7. References

Aizawa, A., Kohlhase, M., and Ounis, I. (2013). NTCIR-10 Math Pilot Task Overview. In *Proceedings of NTCIR-10*.

Alexeeva, M., Sharp, R., Valenzuela-Escárcega, M. A., Kadowaki, J., Pyarelal, A., and Morrison, C. (2020). MathAlign: Linking Formula Identifiers to their Contextual Natural Language Descriptions. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, 2204–2212.

Asakura, T., Greiner-Petter, A., Aizawa, A., and Miyao, Y. (2020). Towards Grounding of Formulae. In *Proceedings of the First Workshop on Scholarly Document Processing (SDP 2020)*, 138–147.

Asakura, T., Miyao, Y., Aizawa, A., and Kohlhase, M. (2021). MioGatto: A Math Identifier-oriented Grounding Annotation Tool. In *13th MathUI Workshop at 14th Conference on Intelligent Computer Mathematics (MathUI 2021)*.

Ausbrooks, R. et al. (2014). Mathematical Markup Language (MathML) 3.0 Specification.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*.

Ginev, D. (2020). arXMLiv:2020 dataset, an HTML5 conversion of arXiv.org. SIGMathLing. `https://sigmathling.kwarc.info/resources/arxmliv/`.

Ginev, D., Stamerjohanns, H., Miller, B. R., and Kohlhase, M. (2011). The LaTeXML Daemon: Editable Math on the Collaborative Web. In *Intelligent Computer Mathematics*.

Miller, B. (2018). LaTeXML *The Manual*—A LaTeX to XML/HTML/MathML Converter, Version 0.8.3.

Oberle, B. (2018). SACR: A Drag-and-Drop Based Tool for Coreference Annotation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Simeone, O. (2018). A Very Brief Introduction to Machine Learning with Applications to Communication Systems. *IEEE Transactions on Cognitive Communications and Networking*.