SemEval 2022 Task 12: Symlink Linking Mathematical Symbols to their Descriptions

[1]Viet Dac Lai, [1]Amir Pouran Ben Veyseh, [2]Franck Dernoncourt, [1]Thien Huu Nguyen
[1]Dept. of Computer and Information Science, University of Oregon, Eugene, Oregon, USA
{*vietl,apouranb,thien*}*@cs.uoregon.edu*
[2]Adobe Research, Seattle, Washington, USA
franck.dernoncourt@adobe.com

Abstract

We describe Symlink, a SemEval shared task of extracting mathematical symbols and their descriptions from LaTeX source of scientific documents. This is a new task in SemEval 2022, which attracted 180 individual registrations and 59 final submissions from 7 participant teams. We expect the data developed for this task and the findings reported to be valuable for the scientific knowledge extraction and automated knowledge base construction communities. The data used in this task is publicly accessible at https://github.com/nlp-uoregon/symlink.

1 Introduction

The exponential growth of published articles may exceeds many readers' ability to keep track of the development of their field of interest. Hence, automatic reading comprehension of scientific documents has attracted the attention of researchers across various domains such as Drug Discovery, Knowledge Base Construction, and Natural Language Processing. A crucial aspect of understanding scientific literature is understanding terminologies and formulae because they offer an explicit and precise interface to present the relation between scientific concepts (Schubotz et al., 2018). As such, a reading comprehension machine needs to (i) identify their descriptions and formulae, (ii) segment them into primitive terms and symbols, and (iii) link the associated terms and corresponding symbols.

Working with mathematical formulae is arduous due to two fundamental reasons. First, common text encodings such as ASCII and Unicode do not fully support typing mathematical symbols. As a result, complex mathematical formulae are rarely written using either ASCII or Unicode. Rather, a higher level encoding (or typesetting) is often used to encode the content of scientific documents, in particular LaTeX. Second, most scientific documents are stored in one of two forms: photos or Portable Document Format (PDF). Scientific documents that were published prior to the graphical computer era are printed and now scanned and distributed as photos. Nowadays, scientific documents are often composed in some text editors or word processing software, then exported and shared a PDF file. Unfortunately, analyzing textual information in photo images or PDF files is extremely difficult, and most of the natural language processing tools are not developed to handle this format. As such, to facilitate the understanding of scientific literature, documents should be stored using a universal easy-to-process text-like encoding. In this paper, we use LaTeX as the typesetting to facilitate document analysis. Thanks to recent advances in text processing and image recognition, a LaTeX document can often be restored to some extent from either a photo or a PDF file (Deng et al., 2017).

This paper introduces the Symlink shared task for the extraction of mathematical symbols and their descriptions from English scientific documents using their LaTeX source. Figure 1 visualizes an example of the task. This paper also presents an analysis of the results of participant systems on the task. The rest of the paper is organized as follows. Section 2 presents related work in extracting formulae and their related information from scientific documents. Section 3 describes the subtasks of this Symlink shared task. Section 4 presents the data creation process including data sources, preprocessing, annotation guidelines, annotation, and data format. An analysis of the created data set is provided in Section 5. The evaluation method is presented in Section 6, while the descriptions of the submitted systems are presented in Section 7.

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1671 - 1678

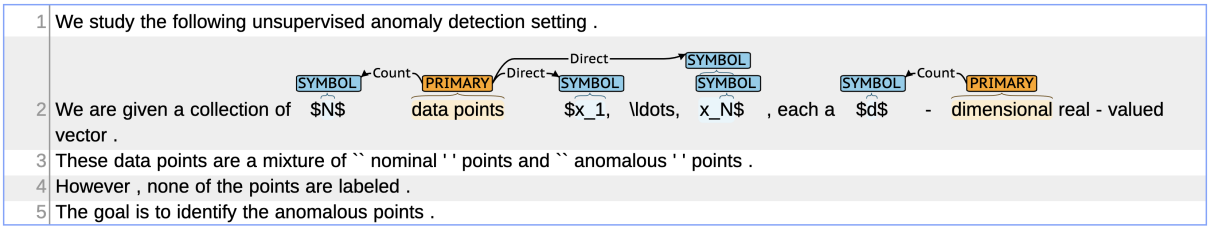

Figure 1: Example of the Symlink tasks.

2 Related Work

Early studies for scientific literature link formulae to Wikipedia page (Nghiem Quoc et al., 2010; Kristianto et al., 2016). Even though this can provide additional information regarding the mathematical expression, a reader might find it harder to understand the Wikipedia page as it is presented in many unrelated forms. Linking to the description in the same document is more practical (Kristianto et al., 2014; Alexeeva et al., 2020) as the descriptions are dedicated to the symbols and the context presented in the document.

Previous studies on symbol-description extraction rely on pattern matching (Yokoi et al., 2011; Nghiem Quoc et al., 2010) and rule-based algorithms (Alexeeva et al., 2020). These methods might work for observed patterns with an assumption of close proximity between symbol and description. They may fail to capture distant symbol-description pairs and symbols in very complex structures such as algorithms in computer science literature.

Most of the previous studies have attempted to extract and link at formula level (Nghiem Quoc et al., 2010; Kristianto et al., 2014, 2016). In reality, understanding mathematical formulae requires details of atomic symbols e.g. superscript, subscript, function arguments. We believe that addressing the problem at this fine-grain level is crucial to drive future research toward a better understanding of the complex symbol-description extraction task.

Prior to this shared task, some studies have created datasets for similar tasks (Yokoi et al., 2011; Schubotz et al., 2016; Alexeeva et al., 2020). However, one of them is created for publications written in Japanese (Yokoi et al., 2011), making it nearly impossible to transfer to English literature. While two other datasets (Schubotz et al., 2016; Alexeeva et al., 2020) only annotate small-scale golden datasets for evaluation purposes. As the result, no training data is available for training deep neural network models. In this shared task, we provide a large-scale dataset for English literature that we believe will provide enough supervision for the promising deep neural network-based models.

Definition extraction from scientific document is close to the task presented in SemEval Task 12. The Scientific Document Understanding workshop has hosted the Acronym Extraction and Acronym Disambiguation Shared Tasks, namely *Acronym Extraction and Acronym Disambiguation Shared Tasks*(Veyseh et al., 2021a, 2022). The prior studies in this research direction considers extracting definitions from the text (Spala et al., 2019, 2020; Veyseh et al., 2020), or together with acronyms, and acronyms sense disambiguation (Pouran Ben Veyseh et al., 2020, 2021).

3 Task Description

The ultimate goal of Symlink shared task is to extract pairs of mathematical symbols and descriptions from scientific documents. As such, Symlink shared task is a combination of an entity recognition and an entity linking task.

Given a LaTeX source of a paragraph from a scientific document:

- **Named Entity Recognition:** For each paragraph, identify all spans containing mathematical symbols and terminology descriptions.
- **Relation Extraction:** For each pair of entities, identify the relationships between them if it is available among symbols and descriptions using Coref-Description, Coref-Symbol, Direct, Count relation types.

4 Data Annotation

4.1 Data source

We obtain the documents from `arXiv.org`, a repository for preprint scientific articles due to the broad coverage of subjects in scientific articles published in ArXiv. In particular, ArXiv offers articles in

physics, mathematics, quantitative biology, computer science, quantitative finance, statistics, electrical engineering, and economics. As such, our obtained papers contain a large number of mathematical symbols and equations, allowing a higher yield of extracted symbol-description relations. Among these subjects, we choose five subjects of mathematics, physics, biology, economics, and computer science for annotation.

4.2 Data preparation

ArXiv open-sources the LaTeX version of their articles, when available. In order to make our Symlink dataset open-access to the whole community, we crawled the metadata of these articles and only selected articles under the CC BY license. Once obtained the LaTeX project, we extracted all the paragraphs from the **.tex** files. We filtered out all short paragraphs with less than 50 words and paragraphs without symbols. Since a formula can be composed in multiple ways such as inline formulae (between $ $), displayed formulae (between $$ $$), or using commands e.g. *array*, to keep the original TeX format of the formulae, all of these math objects are masked before tokenization. Then, we used the SciBERT tokenizer (Beltagy et al., 2019) to tokenize the text. The original math object is then restored. As we observed that many papers have nested math objects, we deleted all the nested objects, hence, having non-nested LaTeX data. This is helpful as it makes the LaTeX documents more similar to the ones generated by the PDF-to-LaTeX tools, which do not contain nested objects.

4.3 Taxonomy

To prepare for the annotation, we designed a taxonomy with 3 general entity types and four relation types. In particular, mathematical symbols are annotated under the tag **SYMBOL**, whereas descriptions are tagged under two labels **PRIMARY**, for single standalone definitions, and **ORDERED**, for the description of multiple terms, whose mentions are not separated without creating non-contiguous mentions. Due to the quadratic numbers of combinations of descriptions and complex math expressions, we only tagged an entity if and only if there is a second entity that pairs with the first entity to form a relationship. For relation, we are particularly interested in two main types of relations: **DIRECT**, linking a symbol with its definition, and **COUNT**, linking a description of a concept with a symbol that is the number of instances of the concept. Due to the sheer number of repetitions and coreferences of both descriptions and symbols, we also annotated **COREF-SYMBOL** relation, linking co-referred symbols, and **COREF-DESCRIPTION** relation, linking co-referred descriptions. Detailed annotation guidelines with examples are presented in Appendix A.

4.4 Annotation

We recruited 10 annotators from the crowdsourcing platform `upwork.com` to annotate scientific papers in the five mentioned domains (each subject was annotated by two annotators). The annotators are explicitly selected based on their demonstrated experiences in reading and writing scientific documents in their expertise field(e.g., holding an M.S. or Ph.D. degree). Detailed annotation guidelines with many examples and explanations are provided to train the annotators. Overall, we annotated 102 papers, accounting for 3,690 paragraphs, and 595K tokens. Our annotators for each domain co-annotate the documents in their domain and achieve Cohen's Kappa scores of (averaged) 0.79. This inter-agreement score thus indicates substantial agreements between our annotators. Eventually, the annotators engage in discussions to resolve any conflict to produce a final consolidated version of our Symlink dataset.

4.5 Data Format

The participants are provided with preprocessed in JSON format. Each paragraph is stored in a JSON object with its id, topic, original LaTeX source, set of entities, and set of relations. An example of the data object is presented in Figure 2.

5 Data Analysis

Table 1 presents the statistics for the dataset including the number of articles, distribution of entities, and distribution of the relations. Overall, our dataset offers more than 31K entities, 20K pairs of relations, which is one order of magnitude larger than existing datasets for a similar task.

Figure 3 presents the distribution of the span lengths of both symbols and descriptions of up to 15 tokens. As can be seen from the figure, the majority of entities have a length of 1-3 tokens. However, overall, the span lengths of both symbols and descriptions vary significantly from 1 up to 47 tokens (note that Figure 3 only illustrates the spans

```
{
  "id": "1503.01158v2...",
  "phase": "test",
  "topic": "cs.ai",
  "document": "1503.01158v2...",
  "paragraph": "paragraph_48",
  "text": "... with a covariance
  matrix of $I$ ; that is , ...",
  "entity": {
   "T1": {
    "eid": "T1",
    "label": "SYMBOL",
    "start": 325,
    "end": 326,
    "text": "I"
   },
   "T2": {
    "eid": "T2",
    "label": "PRIMARY",
    "start": 303,
    "end": 320,
    "text": "covariance matrix"
   }
  },
  "relation": {
   "R1": {
    "rid": "R1",
    "label": "Direct",
    "arg0": "T2",
    "arg1": "T1"
   }
  }
}
```


Figure 2: An example of a paragraph in Symlink dataset.

with up to 15 tokens). This demonstrates a key challenge of the Symbol-Description Linking task in this paper where symbols and descriptions with long spans might introduce confusion for extraction models.

To further understand the dataset, we present the distances between the entities and relations annotated in Symlink by different relation types in Figure 4. The distributions can be grouped into two categories. The first category involves the symbol-description relations while the second group involves the coreference relations. The distributions of symbol-description relations have long tails, indicating that symbols and descriptions tend to appear in close proximity. On the other hand, the distributions of coreference relations are quite flat, suggesting that the coreference relations appear in both short and long distances.

Table 1: Statistics and label distribution of the Symlink dataset. *The texts are tokenized by SciBERT.

	Train	Dev	Test	Total
Statistics				
#Documents	91	6	5	102
#Paragraphs	3,120	270	300	3,690
#Sentences	25,070	1,765	2,286	29,121
#Tokens*	522K	35K	38K	595K
Entity types				
#SYMBOL	18,547	1,504	1,864	21,915
#PRIMARY	7,953	678	907	9,538
#ORDERED	14	3	1	18
Relation types				
#Direct	8,200	731	867	9,798
#Count	1,484	17	221	1,722
#Coref-Symbol	6,821	759	690	8,270
#Coref-Description	612	97	154	863

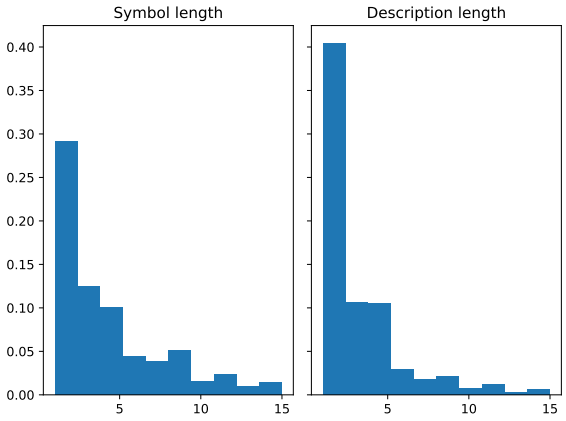

Figure 3: Length of symbols and descriptions in Symlink

6 Evaluation

The results are evaluated separately for the Named Entity Recognition (NER) task and the Relation Extraction (RE) task. For NER, we use the entity-based partial/type from SemEval 2013 Task 9.1. For RE, we use standard precision, recall, F-score metrics. Relations output by the participating system is correct if the prediction label strictly matches the gold standard.

During the 21-day evaluation period (January 10 through 31, 2022), 7 CodaLab users submitted a total of 59 submissions with 37 submissions passing the validation and being scored. Given the complexity of the task, we allow unlimited submissions

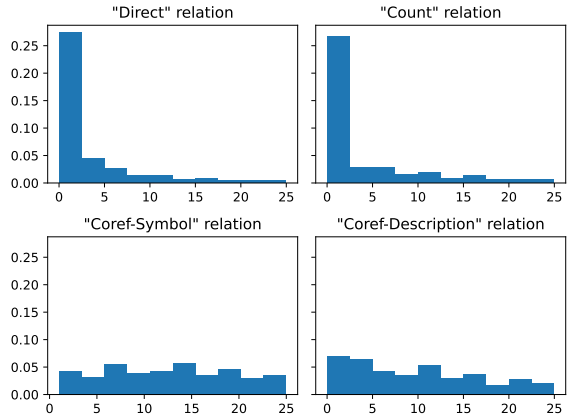

Figure 4: Distribution of distances between entities in Symlink by relation type.

during the evaluation. As such the top submitter tried up to 18 times.

Table 2 shows the performances of the successful submissions. Asterisk denotes teams with system descriptions submitted for review. Among the participated teams, 6 teams performs both Named Entity Recognition and Relation Extraction subtasks while one team tried the Named Entity Recognition subtask only. Figure 5 presents the timelines of submissions and high scores over the evaluation period.

7 Summary of Participating Systems

The Symlink track at SemEval-2022 received 4 system description paper submissions presented in Table 2. Overall, all submitted systems are based on BERT architecture (Devlin et al., 2019). Among those, two out of four systems use SciBERT (Beltagy et al., 2019), while two remaining systems use other variants of BERT such as original BERT (Devlin et al., 2019) and mBERT (Devlin et al., 2019).

7.1 System Specifics

Lee and Na (2022) (JBNU-CCLab) achieved their state-of-the-art performance using SciBERT (Beltagy et al., 2019). Their entity model consists of an MRC-based model (Li et al., 2020), simplifying the tasks as binary classification problems whether span is valid using entity type information as input features. They proposed a simple rule-based Symbol Tokenizer to predict accurately the complex symbols appearing in scientific documents. The relation model exploits entity span information and entity type information as input features using typed entity marker. Additionally, the paper exploited many regularization techniques to improve the model performance such as regularized dropout (Wu et al., 2021) and representational collapse prevention (Aghajanyan et al., 2020) and traditional ensemble techniques.

Popovic and Laurito (2022) (AIFB-WebScience) proposed an end-to-end joint entity and relation extraction approach based on transformer-based language models. Unlike traditional entity and relation extraction methods, which perform the task in sequence, this system incorporates information from relation extraction into entity extraction. As such, the system can be trained even on partially annotated datasets where only a subset of all valid entity spans is annotated.

Ping and Chi (2022) (AN(L)P) participated in the Entity Extraction only. They finetuned a BERT-large model (Devlin et al., 2019) for each domain. For cs.ai domain, they used data from cs.ai only, whereas, for the other domain, they augmented the in-domain data with the data from cs.ai.

der Goot (2022) (MaChAmp) proposed to pretrain a language model and re-finetune after multi-task learning for a pre-defined set of semantically focused NLP tasks. They trained a multi-task model for all text-based SemEval tasks that include annotation on the word, sentence, or paragraph level. They compared the performance with models using mBERT (Devlin et al., 2019). The pretrained multi-task embedding showed a consistent improvement across many tasks against the mBERT embedding.

7.2 Symbol tokenizer and detection

In this shared task, the uniqueness of the task is detecting mathematical symbol span. Symbol span in LaTeX source is comprised of both human language and machine language, i.e. LaTeX language. Further, mathematical formulae in LaTeX sources are written in both linear and hierarchical manners. Therefore, a system must consider not only human language modeling but also a highly systematic syntax system of LaTeX source. As such, fundamental tasks such as tokenization is a huge contributor to the robustness of the model.

Among four submitted systems, MaChAmp (der Goot, 2022) and AN(L)P (Ping and Chi, 2022) teams used the default tokenizer from either BERT or mBERT, which are not designed for scientific documents. Consequently, they are unable to correctly segment the mathematic source, hence, they

Table 2: Results for each team/user, ordered by F1-score on Relation Extraction. Team with $*$ submitted their system description paper to SemEval 2022.

Team	Variant	Entity		Relation		
		F1 (partial)	F1(type)	Precision	Recall	F-score
JBNU-CCLab*	Base	47.61	47.70	32.09	38.56	35.03
	+RDrop	47.61	47.70	33.40	38.66	35.84
	+R3F	47.61	47.70	33.77	38.56	36.00
	+R3F,Ensemble	**47.61**	**47.70**	38.20	**36.23**	**37.19**
ZQ	-	39.39	39.51	**57.25**	23.29	33.11
AIFB-WebScience*	Max/Original	37.83	37.88	45.80	20.96	28.66
	Mean/Original	41.21	41.23	42.25	26.55	32.28
	Max/LaTex2Text	38.33	38.38	46.09	21.64	29.45
	Mean/LaTex2Text	34.53	34.64	47.02	18.20	26.24
LingZing	-	33.87	33.93	13.45	10.92	12.05
MaChAmp*	Single mBERT	-	-	-	-	2.67
	Multi RemBERT	25.17	25.25	13.11	5.17	7.42
iyerke	-	6.67	6.46	0.10	0.62	0.17
AN(L)P*	-	-	16.30	-	-	-

achieved the lowest Named Entity Recognition performance. Whereas AIFB-WebScience (Popovic and Laurito, 2022) and JBNU-CCLab (Lee and Na, 2022) achieved much higher performances thanks to SciBERT tokenizer because it is trained on scientific literature. However, the SciBERT tokenizer is far from perfect such that JBNU-CCLab further proposed to tokenize the mathematical formulae using a customized rule-based tokenizer based on capital letters, numbers, and special characters(e.g. %, $, {, }). Hence, they achieved state-of-the-art performance on both NER and RE subtasks.

8 Conclusion

In this paper, we present the task description, the data annotation, the evaluation, the results, and the descriptions of four submitted systems for Symlink at SemEval 2022. The Symlink shared task is challenging given the complexity of the LaTeX source and partly due to the difference of the domains involved in the data. In this shared task, it is hard to separate the NER and RE subtasks due to their constraints.

The submitted systems employed variants of contextualized embedding BERT for encoding the text. In general, the task can be formatted into similar sequence labeling and relation extraction task. However, special treatments are needed to process LaTeX sources. For instance, a LaTeX-source-trained tokenizer or a customized tokenizer is essential to tokenize the text. Some unique characteristics of the dataset have not been investigated such as the syntax of the LaTeX source, and the hierarchical structure of formulae. These suggest future research directions to improve the robustness of the model.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*.

Maria Alexeeva, Rebecca Sharp, Marco A. Valenzuela-Escárcega, Jennifer Kadowaki, Adarsh Pyarelal, and Clayton Morrison. 2020. MathAlign: Linking formula identifiers to their contextual natural language descriptions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2204–2212, Marseille, France. European Language Resources Association.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M Rush. 2017. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning*, pages 980–989. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rob Van der Goot. 2022. Machamp at semeval-2022 tasks 2, 3, 4, 6, 10, 11, and 12: Multi-task multilingual learning for a pre-selected set of semantic datasets. In *OpenReview*.

Giovanni Yoko Kristianto, Akiko Aizawa, et al. 2014. Extracting textual descriptions of mathematical expressions in scientific papers. *D-Lib Magazine*, 20(11):9.

Giovanni Yoko Kristianto, Goran Topić, and Akiko Aizawa. 2016. Entity linking for mathematical expressions in scientific documents. In *International Conference on Asian Digital Libraries*, pages 144–149. Springer.

Sung-Min Lee and Seung-Hoon Na. 2022. Jbnu-cclab at semeval-2022 task 12: Fusing maximum entity information for linking mathematical symbols to their descriptions. In *OpenReview*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Minh Nghiem Quoc, Keisuke Yokoi, Yuichiroh Matsubayashi, and Akiko Aizawa. 2010. Mining coreference relations between formulas and text using Wikipedia. In *Proceedings of the Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, pages 69–74, Beijing, China. Coling 2010 Organizing Committee.

Annie Ping and Ethan Chi. 2022. Team an(l)p at semeval-2022 task 12: Building a lightweight symbol recognition system. In *OpenReview*.

Nicholas Popovic and Walter Laurito. 2022. Aifb-webscience at semeval-2022 task 12: Relation extraction firstusing relation extraction to identify entities. In *OpenReview*.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Walter Chang, and Thien Huu Nguyen. 2021. MadDog: A web-based system for acronym identification and disambiguation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 160–167, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Quan Hung Tran, and Thien Huu Nguyen. 2020. What does this acronym mean? introducing a new dataset for acronym identification and disambiguation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3285–3301, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Moritz Schubotz, André Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard S Cohl, and Bela Gipp. 2018. Improving the representation and conversion of mathematical formulae by considering their textual context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 233–242.

Moritz Schubotz, Alexey Grigorev, Marcus Leich, Howard S Cohl, Norman Meuschke, Bela Gipp, Abdou S Youssef, and Volker Markl. 2016. Semantification of identifiers in mathematics for better math information retrieval. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 135–144.

Sasha Spala, Nicholas Miller, Franck Dernoncourt, and Carl Dockhorn. 2020. SemEval-2020 task 6: Definition extraction from free text with the DEFT corpus. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 336–345, Barcelona (online). International Committee for Computational Linguistics.

Sasha Spala, Nicholas A. Miller, Yiming Yang, Franck Dernoncourt, and Carl Dockhorn. 2019. DEFT: A corpus for definition extraction in free- and semi-structured text. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 124–131, Florence, Italy. Association for Computational Linguistics.

Amir Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Nguyen. 2020. A joint model for definition extraction with syntactic connection and semantic consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9098–9105.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Thien Huu Nguyen, Walter Chang, and Leo Anthony Celi. 2021a. Acronym identification and disambiguation shared tasks for scientific document understanding. In (Veyseh et al., 2021b).

Amir Pouran Ben Veyseh, Franck Dernoncourt, Thien Huu Nguyen, Walter Chang, and Leo Anthony Celi, editors. 2021b. *Proceedings of the AAAI 2021 Scientific Document Understanding Workshop*, number 2831 in CEUR Workshop Proceedings. Aachen.

Amir Pouran Ben Veyseh, Nicole Meister, Viet Dac Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Acronym extraction and acronym disambiguation shared tasks at the scientific document understanding

workshop 2022. In *AAAI 2022 Workshop on Scientific Document Understanding*, CEUR Workshop Proceedings, Aachen.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34.

Keisuke Yokoi, Minh-Quoc Nghiem, Yuichiroh Matsubayashi, and Akiko Aizawa. 2011. Contextual analysis of mathematical expressions for advanced mathematical search. *Polibits*, (43):81–86.

A Annotation guidelines

This section summarizes some rules that we use to make our annotations more consistent.

Description tagging: A description is usually a noun or a noun phrase that expresses a concept. These are the overall rules for entity annotations:

- We only tag a description if the corresponding symbol presents in the text.
- A description usually is a noun or a noun phrase. Sometimes, a verb, an adverb, or an adjective describes an operation, it is also considered a description.
- Descriptions should be short but it must cover the elements in the corresponding symbol, esp. in case of complex symbols, such as superscript, subscript, arguments, and limits.

Symbol tagging: A mathematical symbol can present an operand, an operator, an expression, or combination of these.

- An atomic symbol in PDF format has to be a character, that means, if we have Y hat, neither Y nor hat is considered an atomic symbol, instead "Y hat" is a symbol. In latex format, \hat{Y} should be annotated.
- A complex symbol is a combination of multiple symbols and brackets, for example: "P(x)", "Wx"
- An annotated symbol has to be a complete symbol e.g. "P(x)" is good, "P(x" is not because of lacking the closing parenthesis.
- A complex formula can be segmented into atomic symbols, we will annotate at all levels of the complex symbol as long as there are appropriate descriptions available.

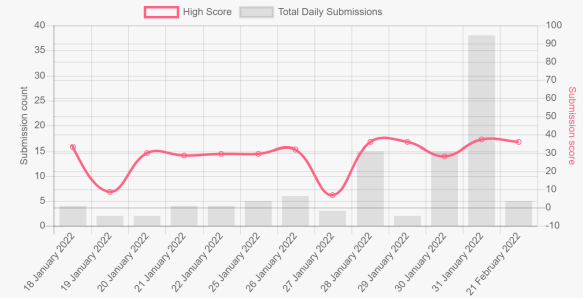

Figure 5: Submission counts and top performances during the evaluation period. The submission score is the F1-score of the RE task.

Relation annotation:

- Every annotated symbol/description has to have at least one relation linking to its description/symbol.
- If there are multiple mentions of a single symbol/description, use coreference relation to link them. A direct relation or a count relation is used to link the closet pair of symbol and description.

B Timeline of submissions

Figure 5 presents the number of submissions over the evaluation of the task.