Symlink: A New Dataset for Scientific Symbol-Description Linking

[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
[2]Adobe Research
{vietl,apouran,thien}@cs.uoregon.edu
franck.dernoncourt@adobe.com

Abstract

Mathematical symbols and descriptions appear in various forms across document section boundaries without explicit markup. Linking mathematical symbols and their descriptions has been conducted for a decade; however, the current state of research is distant toward real application. In this paper, we present a new large-scale dataset that emphasizes extracting symbols and descriptions in scientific documents. Symlink annotates scientific papers of 5 different domains (i.e., computer science, biology, physics, mathematics, and economics). Our experiments on Symlink demonstrate the challenges of the symbol-description linking task for existing models and call for further research effort in this area. We will publicly release Symlink to facilitate future research.

1 Introduction

The exponential growth of published articles exceeds any individual ability to keep track of the development of the field. 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 pivotal angle of understanding scientific literature is understanding their terminologies and the corresponding mathematical formulae because the terminologies and formulae offer an explicit, precise interface to present the relation between scientific concepts [11]. As such, in order to understand the scientific concepts, a reading comprehension machine needs to (i) identify the scientific concept descriptions and their formulae, (ii) segment them into primitive terms and symbols, and (iii) link the associated terms and the corresponding symbols.

1

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, it is impossible to write complex mathematical formulae in plain text using either ASCII or Unicode. As such, a higher level encoding/typesetting is essential to store the content of scientific documents, such as LaTex. Second, most of the 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 as photos. Since the development of graphical computers, scientific documents have been composed and exported to PDF format. Unfortunately, analyzing textual information in photo images or PDF files is extremely difficult, and most of the NLP tools are not developed to handle this type of data. As such, to facilitate the understanding of scientific literature, the documents should be stored using a universal easy-to-process text-like encoding. In this paper, we propose to use LaTeX as the base typesetting to enable the whole analyzing pipeline of scientific documents.

LaTeX is a high-quality typesetting system that is based on the "**what you see is what you mean**" concept, meaning the writers just need to focus on the content of the documents; the computer will take care of the formatting. LaTeX is widely used in producing scientific documents, books, and many other publishing products. LaTeX is a powerful typesetting that it not only produces beautiful documents, but also allows writers to quickly compose complicated document objects such as mathematical symbols and formulae, tables, and graphs. Once a LaTeX document is completed, a LaTeX compiler produces a PDF file with magnificent formatting and those complex document objects. Thanks to recent advances in text processing and image recognition, a LaTeX document can be restored at a high quality from either a photo or a PDF file [3].

In this paper, we focus on extracting mathematical symbols and their corresponding textual descriptions from the LaTeX documents. In particular, given a paragraph as shown in Figure 1, multiple symbols are described within the span of the second sentence, such as the number of data points N, individual data points $x_1, and x_N$, and vector dimension d. Fortunately, the descriptions of these symbols are expressed within the scope of the mentioned sentence. As such, the goal of this work is to link all the symbols and their corresponding descriptions within the context.

Many studies have attempted the problem in various ways. Mathematical symbols can be recognized and linked to Wikipedia pages [8, 6]. Symbols and descriptions in close proximity can be linked [5, 1] because the descriptions are dedicated to the symbols and the context presented in the document. Prior studies attempted to link the description to expressions at formula-level [8, 5, 6]. However, it is essential to understand mathematical expression at a more primitive level, e.g. symbol level. Our work focuses on this fine-grained level annotation for mathematical symbol/description extraction and linking to drive future research toward a better understanding of scientific documents. Moreover, since the prior methods are based on pattern matching and rule-based
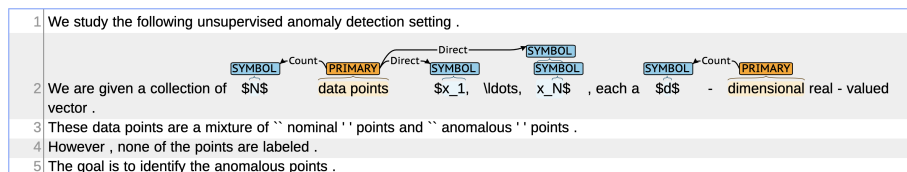
Figure 1: An example of the LaTex document of a scientific article with annotated relations.

algorithms, a few small-scale datasets are created for evaluation purpose only on English literature. As the result, it is impossible to train and evaluate advanced deep learning-based models on these small datasets. This paper presents a large-scale dataset for Symbol Description Linking task for English scientific literature.

The contribution of this paper is three-fold:

- We propose to use LaTex as the intermediate format for scientific documents.

- We create the first large-scale dataset for Symbol Description Linking problem.

- We evaluate the state-of-the-art joint entity-relation extraction models on the new dataset.

2 Data Annotation

We present the annotation process to create a large-scale dataset for Symbol-Description linking in scientific documents.

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.

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 crawl the metadata of these articles and only selected articles under the CC BY license. Once obtained the LaTeX project, we extract all the paragraphs from the **.tex** files. We filter 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 use the tokenizer from SciBERT [2] to tokenize the text. The original math object is then restored. As we observe that many papers have nested math objects, we delete 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 does not contain nested objects.

Annotation taxonomy: To prepare for the annotation, we design 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 massive amount of combinations of descriptions and complex math expressions, we only tag an entity if and only if there is a second entity that pairs with the first entity to form a relation. 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 annotate **COREF-SYMBOL** relation, linking corefered symbols, and **COREF-DESCRIPTION** relation, linking corefered descriptions. Detailed annotation guildlines with examples are presented in Appendix A.

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). Since Upwork allows freelancers to submit a resume for each job application, we explicitly select annotators that have demonstrated experience in reading and writing scientific papers in their corresponding domains (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 101 papers, accounting for 5,719 sentences, and 330K 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 are engaged in discussions to resolve any conflict to produce a final consolidated version of our Symlink dataset.

3   Dataset 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 22K entities, 10K pairs of description and symbol, which is one order of magnitude larger than existing datasets for the similar task.

Figure 2 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

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

	Train	Dev	Test	Total
Statistics				
#Documents	91	6	5	102
#Paragraphs	2,097	201	213	2,511
#Sentences	4,794	412	513	5,719
#Tokens*	283,491	23,036	28,395	334,922
Entity types				
#SYMBOL	11,404	1,199	1,231	13,834
#PRIMARY	7,095	626	762	8,483
#ORDERED	13	3	1	17
Relation types				
#Direct	7,186	703	673	8,562
#Coref-Symbol	2,658	416	302	3,376
#Coref-Description	243	55	51	349
#Count	1,010	17	152	1,179

entities has 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 2 only illustrates the spans 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 3. The distributions can be group in 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 appears in a close proximity. On the other hand, the distributions of coreference relations are quite flat, suggesting that coreference relation relationship appears in both short and long distances.

4 Experiments

To assess the complexity of the Symbol Description Linking task in Symlink, we evaluate the performance of the state-of-the-art deep learning models for ED. In particular, the following models are tested:

- **OneIE** [7]: OneIE is one of the state-of-the-art models in information extraction (IE) that is designed to jointly perform four tasks: entity mention detection, relation extraction, event detection, and argument role labeling. We adapt OneIE to our work using two tasks, i.e., entity mention
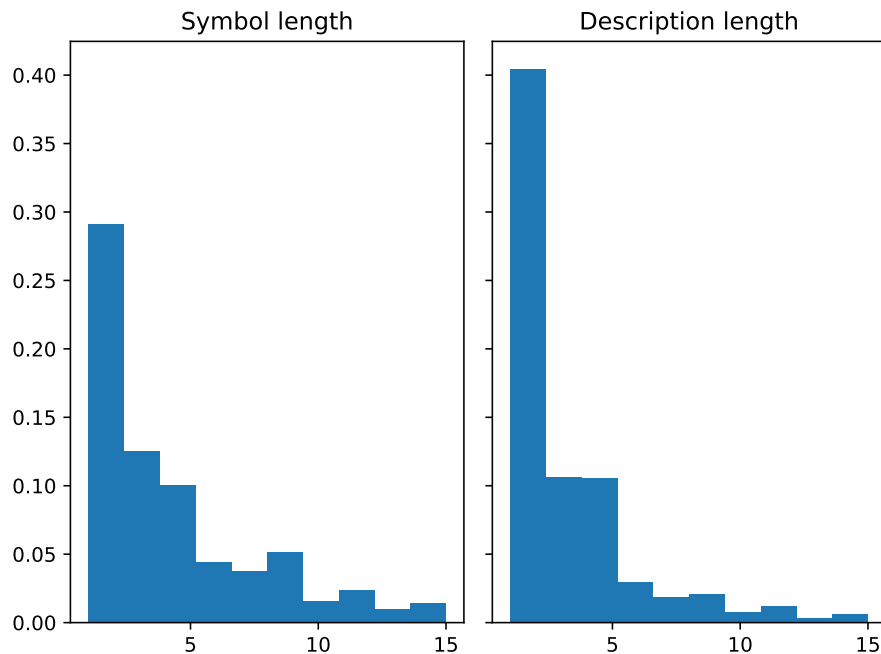
Figure 2: Length of symbols and descriptions in Symlink

detection and relation extraction. The model first detects the spans of entity mentions in the text. The detected spans are then paired to jointly predict entity types and relations.

- **FourIE** [9]: FourIE is another state-of-the-art models for IE to simultaneously solves the four IE tasks in a single model. In addition, FourIE captures the inter-dependencies between prediction instances of different tasks by letting their representations interact with each other, hence, enhancing the predictions of the model. Similar to OneIE, we retain the two modules that perform entity mention detection and relation extraction, thus turning the model into a joint model for entity-relation extraction. This model is the best model on the ACE++ dataset [13].

- **Jerex** [4] is a joint model for entity-level relation extraction from documents. Jerex introduces a multi-task approach that builds upon coreference resolution and gathers relevant signals via multi-instance learning with multi-level representations, combining global entity and local mention information. The model achieves state-of-the-art relation extraction results on the DocRED dataset [14].

Hyper-parameters: As the texts in this dataset are obtained from scientific documents, we use the *allenai/scibert_scivocab_cased* version of SciBERT in
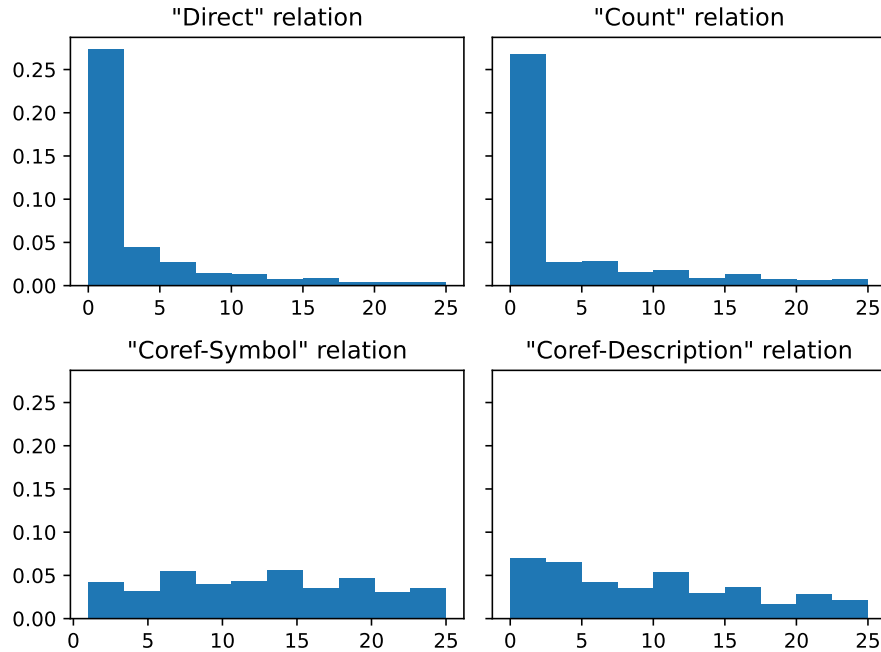
Figure 3: Distribution of distances between entities in Symlink by relation type.

all the considered models [2]. The Stanza toolkit [10][1] is employed for sentence splitting and dependency parsing. The hyper-parameters of the trained models are tuned over the development set based the suggested values from the original papers.

Results: Table 2 reports the performance of the models on the development and test sets of Symlink over two tasks: Entity Mention Detection (Entity) and Relation Extraction (Relation). Among the examined models, OneIE offers the best performance with F1 scores of 54.8% for entity recognition and 32.6% for relation extraction. For comparison, we also include the reported performances of these models on the DocRED and ACE++ datasets [9, 14]. As can be seen, the performance of the current information extraction models on Symlink is significantly and substantially worse than those on the DocRED and ACE++ datasets. For the entity recognition task, the F1-score difference between Symlink and DocRED for Jerex is 35%, and those between Symlink and ACE++ for OneIE is 35%. Due to the low performance of entity recognition, the relation extraction performance hurts sharply by up to 36% between Symlink and ACE++ of the FourIE model. This suggests that the Symbol-Description Linking task in Symlink is a challenging task and more research effort is necessary to boost the performance of the models for this topic.

---

[1]https://stanfordnlp.github.io/stanza/

Table 2: Performance of the state-of-the-art models on Symlink dataset with references of performance on DocRED and ACE++ datasets. The numbers in the right most column show the performances of Jerex on DocRED dataset and OneIE and FourIE on ACE++ dataset. P, R, F stand for precision, recall, f1-score, respectively.

Model	Task	Dev			Test			F
		P	R	F	P	R	F	
Jerex	NER	**82.0**	60.8	69.8	**60.9**	35.7	45.0	80.1
	RE	**72.3**	44.5	55.1	46.9	20.6	28.7	40.4
OneIE	NER	76.9	**69.5**	**73.0**	58.8	51.3	**54.8**	89.6
	RE	61.4	**54.2**	**57.5**	34.2	**31.1**	**32.6**	58.6
FourIE	NER	75.3	68.4	71.7	55.1	**52.2**	53.6	91.1
	RE	59.7	50.5	54.7	28.7	26.6	27.6	63.6

5   Related Work

Early studies for scientific literature link formulae to Wikipedia page [8, 6]. 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 [5, 1] 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 [15, 8] and rule-based algorithms [1]. 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 [8, 5, 6]. In reality, understanding mathematical formulae requires details of atomic symbols e.g. superscript, subscript, function arguments. We believe that address the problem at this fine-grain level is crucial to drive future research toward a better understanding of the complex symbol-description extraction task.

There have been some datasets created Symbol-Description Linking [15]. However, one of them is created for publications written in Japanese [15], making it nearly impossible to transfer to English literature. Moreover, two other datasets [12, 1] only annotate a small-scale golden data for evaluation purpose. As the result, no training data is available for training deep neural network models. In this paper, we provide a large-scale dataset for English literature that we believe will provide enough supervision for the promising deep neural network-based models.

6 Conclusion

We present Symlink, the first large-scale dataset for Symbol-Description Linking on scientific documents. Our experiments demonstrate the poor performance of current information extraction and relation extraction models on the Symlink dataset. This dataset also standardizes the Symbol-Description Linking task, hence facilitating the future research in this topic. In the future, we plan to enlarge the dataset to include more entity and relation types for scientific documents.

References

[1] ALEXEEVA, M., SHARP, R., VALENZUELA-ESCÁRCEGA, M. A., KADOWAKI, J., PYARELAL, A., AND MORRISON, C. MathAlign: Linking formula identifiers to their contextual natural language descriptions. In *Proceedings of the 12th Language Resources and Evaluation Conference* (Marseille, France, May 2020), European Language Resources Association, pp. 2204–2212.

[2] BELTAGY, I., LO, K., AND COHAN, A. 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)* (Hong Kong, China, Nov. 2019), Association for Computational Linguistics, pp. 3615–3620.

[3] DENG, Y., KANERVISTO, A., LING, J., AND RUSH, A. M. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning* (2017), PMLR, pp. 980–989.

[4] EBERTS, M., AND ULGES, A. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Online, Apr. 2021), Association for Computational Linguistics, pp. 3650–3660.

[5] KRISTIANTO, G. Y., AIZAWA, A., ET AL. Extracting textual descriptions of mathematical expressions in scientific papers. *D-Lib Magazine 20*, 11 (2014), 9.

[6] KRISTIANTO, G. Y., TOPIĆ, G., AND AIZAWA, A. Entity linking for mathematical expressions in scientific documents. In *International Conference on Asian Digital Libraries* (2016), Springer, pp. 144–149.

[7] LIN, Y., JI, H., HUANG, F., AND WU, L. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Online, July 2020), Association for Computational Linguistics, pp. 7999–8009.

[8] NGHIEM QUOC, M., YOKOI, K., MATSUBAYASHI, Y., AND AIZAWA, A. 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)* (Beijing, China, Aug. 2010), Coling 2010 Organizing Committee, pp. 69–74.

[9] NGUYEN, M. V., LAI, V., AND NGUYEN, T. H. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Online, June 2021), Association for Computational Linguistics, pp. 27–38.

[10] QI, P., ZHANG, Y., ZHANG, Y., BOLTON, J., AND MANNING, C. D. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (Online, July 2020), Association for Computational Linguistics, pp. 101–108.

[11] SCHUBOTZ, M., GREINER-PETTER, A., SCHARPF, P., MEUSCHKE, N., COHL, H. S., AND GIPP, B. 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* (2018), pp. 233–242.

[12] SCHUBOTZ, M., GRIGOREV, A., LEICH, M., COHL, H. S., MEUSCHKE, N., GIPP, B., YOUSSEF, A. S., AND MARKL, V. 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* (2016), pp. 135–144.

[13] WALKER, C., STRASSEL, S., MEDERO, J., AND MAEDA, K. Ace 2005 multilingual training corpus. In *Technical report, Linguistic Data Consortium* (2006).

[14] YAO, Y., YE, D., LI, P., HAN, X., LIN, Y., LIU, Z., LIU, Z., HUANG, L., ZHOU, J., AND SUN, M. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy, July 2019), Association for Computational Linguistics, pp. 764–777.

[15] YOKOI, K., NGHIEM, M.-Q., MATSUBAYASHI, Y., AND AIZAWA, A. Contextual analysis of mathematical expressions for advanced mathematical search. *Polibits*, 43 (2011), 81–86.

A Annotation guidelines

This section summarizes some rules that we use to make our annotations more consistent. Examples are provided in Table 3.

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.

Table 3: Tag schema for description recognition. The tag name and the entity mentions are highlighted accordingly.

Entity	Symbol	This set of class is divided into base and novel classes C_b and C_n, respectively.
	Primary	Since customers care about fairness and infer sub-proportionally , the price elasticity of demand
	Ordered	is $E = \e+(\e-1) \g \f(M\hat{p}(P))$.
Relations	Direct	The dataset with data-label pairs $D =\{(x\_i, y\_i)\}$ Finally, the monopoly has a constant marginal cost C ¿ 0$
	Coref-Description	The dataset with data-label pairs $D=\{(x_i, y_i)\} \cdots $. The few available labeled data are called support set $ \mathcal{S}=\{(x_i,y_i)\}$
	Coref-Symbol	It chooses price P and output Y to maximize profits $(P - C) \cdot Y$ subject to customers' demand
	Count	$\cdots$ where every class in the task has q test cases .

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 there is "\hat{Y}" in the document, neither "Y" nor "\hat" is considered an atomic symbol, instead "\hat{Y}" is a symbol.

- 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.

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.