Chemical Language Understanding Benchmark

Yunsoo Kim Hyuk Ko Jane Lee Hyun Young Heo Jinyoung Yang Sungsoo Lee Kyu-hwang Lee
LG Chem
{kys.930303, kohyuk, janelee, hyheo11, yang.jy, ssoolee, amine}@lgchem.com

Abstract

In this paper, we introduce the benchmark datasets named CLUB (Chemical Language Understanding Benchmark) to facilitate NLP research in the chemical industry. We have 4 datasets consisted of text and token classification tasks. As far as we have recognized, it is one of the first examples of chemical language understanding benchmark datasets consisted of tasks for both patent and literature articles provided by industrial organization. All the datasets are internally made by chemists from scratch. Finally, we evaluate the datasets on the various language models based on BERT and RoBERTa, and demonstrate the model performs better when the domain of the pre-trained models are closer to chemistry domain. We provide baselines for our benchmark as 0.7818 in average, and we hope this benchmark is used by many researchers in both industry and academia. The CLUB can be downloaded at https://huggingface.co/datasets/bluesky333/chemical_language_understanding_benchmark.

1 Introduction

Transformer is the prevalent network architecture in natural language processing (NLP) (Vaswani *et al.*, 2017). It uses self-attention to capture each word's influence on another in a given text. Leveraging this architecture, recent advances in pre-training language models has reached state-of-the-art performances on many NLP benchmark datasets, including results that surpassed human performance (Wang *et al.*, 2019). Such advancements in language models and NLP technologies can potentially streamline and simplify the labor-intensive work for the literature and patent analysis, which are crucial in the research and development domain.

The benchmark datasets such as GLUE and SuperGLUE played a pivotal role in facilitating the advancement of NLP using language models (Wang *et al.*, 2018 and Wang *et al.*, 2019). This has inspired efforts to create benchmark datasets in the science domain as well (Yu Gu *et al.*, 2020). However, these attempts are limited within the field of biology and medicine.

In chemistry, there are few datasets available, however, as far as we know there are no benchmark datasets that include tasks for both literature articles and patents (Mysore *et al.*, 2019, Friedrich *et al.*, 2020, He *et al.*, 2021). Given the predominant reliance on patents in the chemical industry's research, especially in the early stages of product development, it is important to have datasets with patent documents to enable language models to comprehend the distinctive patent writing style, thereby performing better on tasks with patent documents.

On the other hand, academic literature often serves as the source of information that leads to new ideas for experimentation. Thus, it is critical to build a language model that understands both literature articles and patents and benchmark datasets with texts from both patents and papers for the evaluation.

In this paper, we present Chemical Language Understanding Benchmark (CLUB) to facilitate NLP research in the chemical industry, especially the language model pre-training. CLUB consists of two datasets for patents and two datasets for papers. In terms of tasks, it includes two datasets for token classification such as chemical named entity recognition, and two datasets for text classification such as patent area classification. All these datasets are internally made by chemists. We do not rely on any preexisting publicly available datasets or shared tasks. Finally, we provide the performance of various language models including the ones pre-trained with chemistry literature articles and patents as the baselines for our benchmark datasets.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 404–411

Tasks	Class Group (source corpus)	Sample Type (Number)	Average token length (std)	Class name	Definition	Train	Dev
Text CLS	PETRO-CHEMICAL (Patent)	Paragraph (2,775)	448.19 (403.81)	Household	Patents for products used in household such as PET bottles	436	120
				Construct	Patents for products used in construction such as PVC pipes	77	25
				Automobile	Patents for products used in automobile such as Tires	312	89
				HouseConst	Patents for products used in household and construction	481	93
				IndustConst	Patents for products used in industrial and construction	274	62
				Catalyst	Patents for catalyst used for production	334	94
				Process	Patents for production process of the products	306	72
	RHEOLOGY (Journal)	Sentence (2,017)	55.04 (16.46)	Biodegrad_Poly	biodegradable polymer (plastic material)	553	151
				Poly_Struc	the crystal structure of polymer which is related with mechanical properties	421	105
				Biodgrad_Prop	biodegradable property of polymer	470	97
				Mechanical_Prop	mechanical property of polymer	90	31
				Rheological_Prop	rheological property of polymer which is related with polymer processability	78	19
Token CLS	CATALYST (Patent)	Sentence (4,663)	42.07 (14.59)	Precatalyst	Pre-catalyst form of metallocene catalyst	365	71
				Olefin	Include monomers and comonomers that participate in the synthesis of supported catalyst	947	153
				Solvent	A solvent that creates a reaction environment	1,287	356
				Additive	Additives necessary for the catalyst synthesis reaction include scavengers and cocatalysts.	402	131
				Support	Support material for synthesis	417	83
	BATTERY (Journal)	Sentence (3,750)	40.73 (10.79)	Cathode_Material	Lithium compound used for cathode electrode among the components of lithium ion battery	1,411	402
				Coating_Material	Materials coated for the purpose of improving structural stability and chemical resistance of cathode materials	1,510	359
				Coating_Method	Method for coating the coating material on the surface of the cathode material	409	134

Table 1: CLUB datasets for text and token classification (CLS).

2 Tasks

The CLUB Benchmark is created from scratch to evaluate language models that understand the fields of chemistry and materials science. The benchmark dataset includes two types of tasks: text classification and token classification. To evaluate the representation power of the language model for both patents and literature articles, each task consisted of a dataset created from the patent text and a dataset created from the paper text. Various topics such as polymers, rheology, catalysts, and batteries were selected to evaluate different fields of chemistry and materials science. The detailed composition of the data set is summarized in Table 1.

2.1 Text Classification

Text classification task is to assign a sentence or document to a proper class. In this paper, we present two classification datasets: RHEOLOGY for sentence classification and PETROCHEMICAL for document classification. These datasets comprise corpora from both patents and journal articles with a focus on the topics of polymers, rheology, and overall petrochemicals. Each dataset is available in JSON format with "id", "sentence", and "labels" as keys.

RHEOLOGY sentence classification dataset contains the five groups that represent the polymer structures and properties, especially for biodegradable polymers. It consists of 2,017 sentences collected from the research paper. Each sentence of the RHEOLOGY classification dataset is annotated by experts manually. The detailed explanation of each group is presented in Table 1.

PETROCHEMICAL dataset categorizes patents into seven groups within the petrochemical industry. Each group of patents accounts for important parts of the industry. The petrochemical industry uses **catalysts** to make the final polymer products for different applications such as PET bottles (**household applications**), rubber (**automobile applications**), and PVC plastics (**construction applications**). This production is done on a factory scale, so it has its **production process**. The seven groups consist of 5 applications: 1) household, 2) automobile, 3) construction, 4) household & construction, and 5) automobile & construction. The other two groups are catalysts and processes.

2.2 Token Classification

Token classification, which includes named entity recognition task, identifies tokens belonging to defined classes. Considering our interests, we defined the CATALYST class group and the BATTERY class group as shown in Table 1. We created the named entity recognition benchmark dataset based on these definitions. The labeling was performed by expert researchers with over five years of experience in relevant fields. The labeling was done in IOB format (inside, outside, beginning). The labeled data was then converted into JSON format with "id", "tokens", and "labels" as keys.

We preprocess the token classification datasets to adjust the sentence length to be less than the maximum sequence length. As for named entity recognition, each token has labels, and tokens that come after the maximum sequence length would be discarded. Thus, the model would not be able to learn from those discarded tokens. We minimized this issue by making the distribution of the sequence length more like the gaussian distribution (Appendix A).

CATALYST is a dataset for recognizing materials involved in catalyst synthesis reactions in the full text of patents. Pre-catalyst, additive, olefin, solvent, and supporting material are substances that participate in this reaction, and these are defined as classes. "Pre-catalyst" is the main substance to make the catalyst. "Additives" are added to make the polymer with different characteristics. "Olefin" is the monomer that makes the polymer using the catalyst. "Solvent" is for the polymerization of the monomer to the polymer for the catalyst. "Supporting material" is used to support the catalyst to do the polymerization better as well as more stable.

BATTERY is a dataset for recognizing cathode materials from literature articles related to lithium-ion batteries including all-solid-state batteries. There are four key components of a battery: cathode material, anode material, separator, and electrolyte. "Cathode material" refers to the lithium compound used in the positive electrode of a battery and is the most important element in a battery because it has a decisive effect on the energy density, power output, and cycle life of the battery. This dataset also has "coating material" and "coating method" classes which are material and method to coat the surface of the cathode material.

3 Dataset Statistics

All datasets have been divided into a training set and a development set (also known as the evaluation set), following an 80/20 split ratio.

3.1 PETROCHEMICAL dataset

The PETROCHEMICAL dataset is composed of 2,775 paragraphs. As the dataset is created with titles, abstracts, and claims of patents, so it has the average paragraph length of 448.19 tokens, which is considerably longer than the other three datasets. Also, the standard deviation for the paragraph length is 403.81 tokens, which is also larger than the others. For the seven classes of the dataset, the respective counts of paragraphs are as follows: "Household" – 556, "Construct" – 102, "Automobile" – 401, "HouseConst" – 574, "IndustConst" – 336, "Catalyst" – 428, and "Process" – 378.

3.2 RHEOLOGY dataset

The RHEOLOGY dataset is made up of 2,017 sentences with an average sentence length of 55.03 tokens. The standard deviation of the sentence length is 16.46 tokens. 704 sentences were labeled as "Biodegrad_Poly" class and 526 sentences were labeled as "Poly_Struc". The "Biodegrad_Prop", "Mechanical_Prop", and "Rheological_Prop" classes, which are classes related to material's properties, were labeled with 567, 121, and 97 sentences, respectively.

3.3 CATALYST dataset

The CATALYST dataset consists of 4,663 sentences. The average sentence length is 42.07 tokens with 14.59 tokens for standard deviation. "Solvent" class was labeled the most with 1,643 times, followed by "Olefin" class which as labeled 1,100 times. "Precatalyst", "Addtive", and "Support" were labeled 436, 533, and 500 times, respectively.

3.4 BATTERY dataset

The BATTERY dataset consists of 3,750 sentences, and the average sentence length is 40.73 tokens with 10.79 tokens as standard deviation. The token classification breakdown shows that "Cathode_Material" and "Coating_Material" classes were labeled 1,813 times and 1,869 times, respectively. Meanwhile, the "Coating_Method" class was 543 times.

4 Methods

4.1 Baseline Models

BERT-Base We use the BERT-base weights released on Hugging Face model repository (Devlin *et al.*, 2018). Both **cased** and **uncased** versions of the model are used. We refer to each version as **BERT-cased** and **BERT-uncased** respectively throughout our papers. The model is pre-trained with a corpus made up of BooksCorpus and text parts of English Wikipedia for 1 M steps. The corpus is about 16GB. The pre-training batch size is 256 sequences. This model utilizes a wordpiece vocabulary. The vocab size is 28,894.

BioBERT We use **BioBERT-v1.2** weights released on Hugging Face model repository (Lee *et al.*, 2020). This is a BERT-base-cased model pre-trained with PubMed abstracts from the BERT-base-cased initial checkpoints. It was trained for 200K steps on PubMed abstracts, 270K steps on PubMed Central (PMC) full texts, and another 1 M steps on PubMed abstracts. The pre-training corpus is about 25GB. The pre-training batch size is 192. As a continued pre-trained model, it uses the same vocabulary as the **BERT-base-cased** model.

SciBERT We use **sciBERT-scivocab-uncased** released on Hugging Face model repository (Beltagy *et al.*, 2019). This is a pre-trained BERT model with 1.14 M Semantic Scholar papers, which is comprised of computer science (18%) and biomedical domain (82%). It differs from BioBERT as it is pre-trained from scratch. The papers are full texts and resulting in a corpus size of 20GB. The pre-training batch size and steps are unknown. It has its own wordpiece vocabulary made from the pre-training corpus. The vocabulary has more science terms. The vocab size is 30,990.

RoBERTa We use **RoBERTa-base** model released on Hugging Face model repository (Liu *et al.*, 2019). It is an improvement of BERT model with a larger pre-training dataset and better optimized hyperparameter settings. The model is pre-trained with a 160GB corpus made up of BERT pre-training corpus plus News and Web contents crawled. It is trained for 1 M steps. The pre-training batch size is 256 sequences. The model uses byte pair encoding (BPE) vocabulary, which is different from BERT's wordpiece vocabulary. The vocab size is 50,000.

	Text classification (Accuracy)		Token classification (F1)		
Task	RHEOLOGY	PETRO-CHEMICAL	CATALYST	BATTERY	Average
BERT-cased	0.7970	0.8099	0.6601	0.7532	0.7550
BERT-uncased	0.7921	0.8105	0.6944	0.7571	0.7635
RoBERTa	0.7958	0.7990	0.6899	0.7658	0.7626
BioBERT	0.7978	0.8086	0.7092	0.7636	0.7698
SciBERT	0.7938	0.8045	0.7314	0.7602	0.7724
RoBERTa-PM-M3	0.7983	0.8079	0.7194	**0.7815**	0.7767
RoBERTa-lit	**0.8017**	0.8126	**0.7332**	0.7772	0.7811
RoBERTa-lit-pat	0.7968	**0.8205**	0.7323	0.7777	**0.7818**

Table 2: Performance of the model for the benchmark tasks. The evaluation for the text classification tasks was done using accuracy and the evaluation of the token classification tasks was done using macro-average of F1 scores. The evaluation result is the average of performances over ten runs.

RoBERTa-PM-M3 We use **RoBERTa-base-PM** weights released on Hugging Face model repository (Lewis *et al.*, 2020). It is a **RoBERTa-base** model pre-trained with a text corpus made of 27GB of PubMed abstracts, 60GB of PMC full texts, and 3.3 GB of the Medical Information Mart for Intensive Care (MIMIC-III). The model is trained for 500K steps on the corpus with a batch size of 8,192 sequences. It uses byte-pair encoding vocabulary made from the corpus, so it has a different BPE encoding vocabulary from RoBERTa-base. The vocabulary has more biomedical terms. The vocab size is 50,000.

4.2 Pre-training

For the chemistry pre-training, we gathered a large amount of chemistry patents and literature articles to train two different versions of models.

RoBERTa-lit We use **RoBERTa-PM-M3** weights as the initial checkpoint to pre-train the model with chemistry articles. We collected the abstracts of the articles using Open Academic Graphs and used the chemistry field of study to filter the ones that belong to the chemistry domain (Tang *et al.*, 2008 and Sinha *et al.*, 2015). For the filtered ones, all the abstracts were used as the training corpus. We train the model with the corpus for 1 epoch.

RoBERTa-lit-pat We use **RoBERTa-lit** weights as the initial checkpoint to pre-train the model this time with chemistry patents. We collected the patents using USPTO BulkDownload. We filtered the chemical patents using the CPC code. For the filtered ones, abstracts, claims, and embodiment texts were used as the training corpus together with the **RoBERTa-lit**'s corpus. We train the model with the corpus for 1 epoch.

RoBERTa-lit and **RoBERTa-lit-pat** were pre-trained with NVIDIA V100 GPU and the hyperparameter setting follows the pre-training setup for **RoBERTa-PM-M3**. We also used mixed precision for training. We used the masked language model objective for the pre-training.

We expect that by pre-training the models with chemistry data, the models can learn the chemistry domain knowledge better and thus perform better on the CLUB benchmark.

4.3 Finetuning Language Models

For each dataset, we fine-tuned each models for 10 epochs with a 5e-05 learning rate on a single V100 GPU. We used 0.1 warm-up ratio, and cosine with restarts as the learning scheduler type. The training batch size was 128 and the evaluation batch size was 128. The maximum input length was 256. AdamW was used as the optimizer with a weight decay of 0.01. We used mixed precision for efficient training. We fine-tuned the model for 10 different seed initializations.

4.4 Evaluation

We evaluated all models using the accuracy for text classification tasks and the macro-average F1 score for token classification tasks. We chose the accuracy as the evaluation metric for the text classification due to its interpretability in measuring the effectiveness of the models. For token classification tasks, the use of the IOB scheme, which resulted in the "O" label being the dominant class, limited us from using the evaluation metric as text classification tasks. To provide a more balanced evaluation, we computed the F1 score of each token class excluding the "O" class, and used the macro-average of these F1 scores as the evaluation metric. For both types of tasks, the performance was averaged over ten runs with different seed initializations to reduce variance caused by randomness.

5 Results and Discussion

The performance of each model on the benchmark tasks is shown in Table 2. In general, our RoBERTa-lit-pat model outperformed the other models on average across the tasks. The result of BioBERT models pre-trained with a bio-related corpus was better than that of BERT base models, highlighting the impact of domain specific pre-training. SciBERT model pre-trained with a broad scientific literature articles performed well, especially in CATALYST task, though it still had a lower performance than RoBERTa models pre-trained with chemistry corpus. RoBERTa-PM-M3 model outperformed other models in the BATTERY task, but its overall performance was lower than that of the RoBERTa-lit-pat model.

In the text classification task, RoBERTa-lit model was the best model in the RHEOLOGY task and RoBERTa-lit-pat model score the highest in the PETROCHEMICAL task. This suggests that inclusion of patents in pre-training yields better performance in tasks with patent documents. As the PETROCHEMICAL dataset includes titles, abstracts, and representative claims of patents, the terminology used in the dataset is quite different from the terminology used in other datasets made up of literature articles. This is due to the nature of patents to protect an invention, leading them to be written in a more general manner to encompass a broader patent space.

In the CATALYST task, it was very interesting that RoBERTa-lit model, solely pre-trained on academic papers, showed the best results in the task with patents. This task involved labeling only the embodiment section of the patent. The terminology used in the embodiment part of the patent is closer to academic language than the language used in patent claims. This could explain why a model trained only on articles could perform better in this task.

For the BATTERY task, RoBERTa-PM-M3 model had the best performance, closely followed by RoBERTa-lit-pat model. Notably RoBERTa-lit and RoBERTa-lit-pat models still showed good average performance despite only being pre-trained for one epoch. It is plausible that the performance of RoBERTa-lit-pat improves further with additional training epochs. Due to our GPU infrastructure limitations, we leave this for future work.

6 Conclusion

Chemical Language Understanding Benchmark (CLUB) is the first benchmark in the chemistry industry aimed at chemical language model evaluation with tasks for both patents and journal articles. The introduction of this benchmark is expected to catalyze research in natural language processing, particularly in information extraction, within the chemistry domain.

In the course of establishing baseline performance for the CLUB, we tested existing pre-trained models as well as our novel pre-trained models. Remarkably, the RoBERTa model pre-trained on cheimcal patents and literature articles, reached the highest average score, 0.7818. This performance highlights the advantage of pre-training models with a corpus closely aligned with the target domain.

Our benchmark provides a powerful tool for evaluating language models' learning capacity in the chemistry context. In addition, the tasks in our benchmark can be leveraged to accelerate the literature and patent analysis by automatically extracting information such as new chemical molecules and experiment settings.

Thus, these tasks can be the foundation of an information extraction based expert system. This system would generate structured knowledge from a large volume of papers and patents and help researchers to conduct their experiments on time without falling behind the research trends.

Our benchmark sets the foundation for future advancements in chemical language understanding.

It contributes to the acceleration of scientific discovery in the field by integrating natural language processing into chemical research and development.

Limitations

Because we were doing the manual labeling with experts in the field, we were only limited to two types of tasks: token classification and text classification. We hope to expand the benchmark to include other types of tasks such as summarization, question and answering, and sentence similarity in the future. Sentence similarity for patents is the task we aim to add for the next version because it can be used to find the infringement in patents.

While the CLUB provides a robust benchmark for evaluating language models in the context of chemistry, it is not without its limitations. The present version of CLUB only includes two types of tasks: token classification and text classification. This constraint arises primarily from the manual labeling process which involved domain experts.

However, we aim to extend the benchmark in the future to include a wider range of tasks such as summarization, question answering, and sentence similarity assessments. We are particularly interested in the sentence similarity task for patents as this could be leveraged for identifying potential patent infringements.

Acknowledgements

We express our sincere gratitude to the anonymous reviewers who contributed their valuable time and effort to provide insightful and constructive feedback on this work. Your detailed comments and suggestions have greatly aided us in refining our work. We would also like to extend our thanks to everyone involved in the creation and development of the labeled datasets. This work would not have been possible without your dedication and collaborative efforts. Last but not the least, we are grateful for the continuous support and resources provided by our institutions, which have been fundamental in conducting this research.

References

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv preprint arXiv:1804.07461.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. arXiv preprint arXiv:1905.00537.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, pages 243-246.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. arXiv preprint arXiv:1903.10676.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2021. Chemu 2020: Natural language processing methods are effective for information extraction from chemical patents. *Frontiers in Research Metrics and Analytics*, 6, 654438.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and Mining of Academic Social Networks. 2008. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*. pages 990-998.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So and Jaewoo Kan. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Volume 36, Issue 4, February 2020, pages 1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, Elsa Olivetti. 2019. The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. arXiv preprint arXiv:1905.06939.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. In *ACM Transactions on Computing for Healthcare*, pages 1–23.

A Adjust sentence length

Figure 1. shows the distribution of sentence lengths in the dataset before and after the preprocessing. After adjusting the sentence length, the sequence length distribution follows more of a Gaussian distribution than before. In the case of CATALYST dataset, the number of sentences was reduced from 12,368 to 4,663. However, in the case of BATTERY dataset, there was no change in the number of the sentences. We made this preprocessing to minimize the number of tokens that come after the maximum sequence.

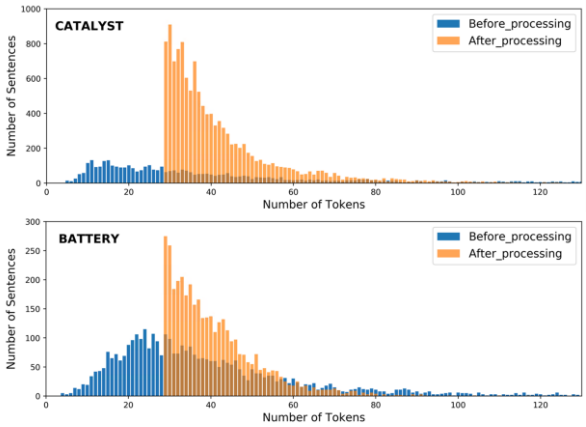

Figure 1. Distribution of sequence length before and after sentence adjustment in token classification task datasets