PolyMinder: A Support System for Entity Annotation and Relation Extraction in Polymer Science Documents

Dinh-Truong Do[1,2*] **Hoang-An Trieu**[1,2*] **Van-Thuy Phi**[2]
Minh Le Nguyen[1] **Yuji Matsumoto**[2]

[1]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{truongdo, antrieu, nguyenml}@jaist.ac.jp
[2]RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
{thuy.phi, yuji.matsumoto}@riken.jp

Abstract

The growing volume of scientific literature in polymer science presents a significant challenge for researchers attempting to extract and annotate domain-specific entities, such as polymer names, material properties, and related information. Manual annotation of these documents is both time-consuming and prone to error due to the complexity of scientific language. To address this, we introduce PolyMinder, an annotation support system designed to assist polymer scientists in extracting and annotating polymer-related entities and their relationships in scientific documents. The system utilizes recent advanced Named Entity Recognition (NER) and Relation Extraction (RE) models tailored to the polymer domain. PolyMinder streamlines the annotation process by providing a web-based interface where users can browse, verify, and refine the extracted information before obtaining the final results. The system's source code is made publicly available to facilitate further research and development in this field. Our system can be accessed through the following URL: https://www.jaist.ac.jp/is/labs/nguyen-lab/systems/polyminder.

1 Introduction

The study of polymers has gained significant momentum in recent years, driving advancements in diverse fields, including materials science, manufacturing, biomedical engineering, and environmental sustainability (Okolie et al., 2023; Sharma et al., 2021). Their versatility and wide-ranging properties make them indispensable in products such as plastics, rubbers, adhesives, and nanomaterials (Mohanty et al., 2022; AlMaadeed et al., 2020). As research in this area continues to grow,

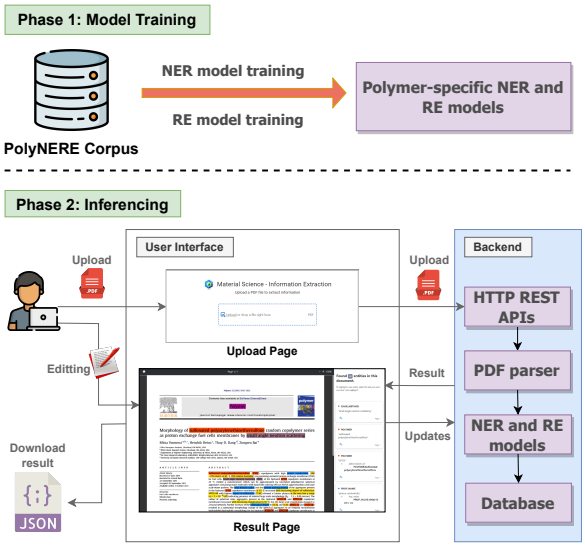


Figure 1: An overview of PolyMinder, showcasing its workflow from model training to entity/relation extraction and user interaction via the web-based interface.

the volume of scientific literature on polymer structures, properties, and applications has grown substantially (Phi et al., 2024). Efficiently organizing and extracting valuable information from this vast corpus is crucial for both research and industry. However, manually extracting and annotating domain-specific entities like polymer names and material properties is challenging (Fagnani et al., 2022). The complexity of scientific language and the specialized expertise required make manual annotation time-consuming and error-prone.

While automated named entity recognition (NER) systems that use advanced neural networks have shown promise in the materials domain, including polymers (Oka et al., 2021; Phi et al., 2024; Cheung et al., 2024), existing solutions often accept only text-based inputs and output results in formats like JSON, lacking intuitive visualizations for users or annotators. They typically do not handle PDF inputs directly, despite PDFs being standard in the scientific community. Moreover, neural network

*Equal contribution

models are not infallible and can produce errors in entity extraction. Current systems lack features that allow annotators to efficiently refine or correct extracted data and parsing errors, necessitating manual adjustments without system support.

To bridge this gap, we present PolyMinder, an automated support system tailored specifically for the polymer domain. PolyMinder aids researchers by automatically identifying and extracting key information from scientific documents—such as polymer names, material properties—enabling the visualization and refinement of extracted data while significantly reducing manual effort. Figure 1 provides an overview of our system. In the training phase, we developed entity and relation extraction models using recent advanced techniques (Li et al., 2022; Zhou et al., 2021), trained on PolyNERE (Phi et al., 2024), a high-quality corpus for named entity recognition and relation extraction, covering a variety of entity and relation types. During inference, users can upload documents to the system, which extracts content from PDF files. The entity and relation extraction models process this content to identify relevant entities and relationships, which are then displayed in an intuitive web interface. Users can review and refine the automated extractions for accuracy before downloading the fully annotated output. In summary, our contributions in this paper are threefold: **(I)** we introduce PolyMinder, an automated support system that extracts and visualizes key polymer-related information from scientific texts, allowing annotators to review and refine the data; **(II)** we develop polymer-specific entity and relation extraction models utilizing state-of-the-art techniques tailored to the polymer science domain; and **(III)** we publicly release PolyMinder's source code[1] to support further research and development.

2 Related Works

Automated information extraction from scientific literature has advanced in fields like biomedical science, chemistry, and materials science (Yang et al., 2022). In polymer science, Oka et al. (2021) developed a system combining deep learning with rule-based methods to extract polymer data from tables in scientific articles. Similarly, Swain and Cole (2016) introduced ChemDataExtractor, a tool that processes chemical data using rule-based and supervised learning approaches. Weston et al. (2019) created MatScholar, which applies named entity recognition to extract key information from materials science literature. In biomedicine, the BENNERD system (Sohrab et al., 2020) targets COVID-19-related entity extraction, while Wadhwa et al. (2021) proposed a system for extracting fabrication knowledge in materials science, identifying key entities and relationships. More recently, Shetty et al. (2023) fine-tuned MaterialsBERT to extract material property records from large polymer corpora, significantly outperforming baseline models like BioBERT (Lee et al., 2020) and ChemBERT (Davronov and Adilova, 2023). Despite these advancements, the models used in these systems are still prone to errors, and current systems provide limited support for efficiently correcting them.

From the perspective of annotation tools, most existing solutions focus on entity recognition and relation extraction in plain text formats (Borisova et al., 2024). Tools like Brat (Stenetorp et al., 2012) and Doccano (Nakayama et al., 2018) offer a visual interface for annotating entities and relations in natural language texts, aiding manual curation. However, these tools lack direct interaction with PDF documents, the standard format for scientific research. PDFAnno (Shindo et al., 2018) addresses some of these limitations by enabling users to annotate entities and relations directly on PDF documents, preserving the original layout. Despite this, PDFAnno can become cluttered when annotating documents with numerous relations, as the visualization of arrows may overwhelm the interface and hinder the annotation process. This highlights the need for a more intuitive web-based interface that offers easy, clear annotation on PDF documents.

3 Method

3.1 Overview

Our proposed system, PolyMinder, is designed to extract polymer-related entities and their corresponding relationships from scientific documents (PDFs), followed by a visualization of these entities and relationships on a web interface for annotators to verify and refine (Figure 1). The process begins with training named entity recognition and relation extraction models using recent advanced methods, specifically ALTOP (Zhou et al., 2021) and W2NER (Li et al., 2022). Once the models are trained, we develop the web application, which includes both the frontend interface and the supporting backend infrastructure.

[1] https://github.com/truongdo619/PolyMinder

3.2 Entity and Relation Extraction Models

Entity extraction serves as the basis for identifying key polymer-related entities within documents. For this task, we employ the W2NER model (Li et al., 2022), which outperforms top-performing models on widely-used benchmark datasets on general, biomedical, and clinical domains. W2NER utilizes a word-to-span alignment approach, which allows it to handle three types of entity mentions (flat, overlapped, and discontinuous) effectively, a common challenge in scientific text. This model is trained on the PolyNERE corpus (Phi et al., 2024), a domain-specific dataset containing 750 polymer abstracts across 14 entity types. All entity types used in our system are summarized in Figure 2.

```
ENTITY
|------ MATERIAL_GROUP
|       |------ POLYMER_GROUP
|       |       |------ POLYMER
|       |       |------ POLYMER_FAMILY
|       |------ INORGANIC
|       |------ ORGANIC
|       |       |------ MONOMER
|       |------ COMPOSITE
|       `------ OTHER_MATERIAL
|------ PROP_NAME
|------ PROP_VALUE
|------ MATERIAL_AMOUNT
|------ CONDITION
|------ CHAR_METHOD
|------ SYN_METHOD
`------ REF_EXP
```

Figure 2: Ontology of material entities in PolyMinder, showing material types. Monomers fall under organic, while polymers may be organic or inorganic.

After identifying the entities, the next step is to establish their relationships. In polymer science, these relationships represent the essential connections between polymer-related entities. For instance, the entity `POLYMER` is related to the `CHAR_METHOD` entity through the `characterized_by` relationship. The details of all entities and their relationships in our system are illustrated in Figure 3. To extract these relationships, we tackle the RE task at the paragraph level, focusing on predicting relationships between entity pairs within the text. We utilize the ATLOP model (Zhou et al., 2021), designed for document-level or paragraph-level RE. By employing transformer-based attention mechanisms, ATLOP effectively captures complex, cross-sentence relationships.

After training, the NER and RE models integrate into PolyMinder's backend, generating JSON outputs. An example of this output is as follows:

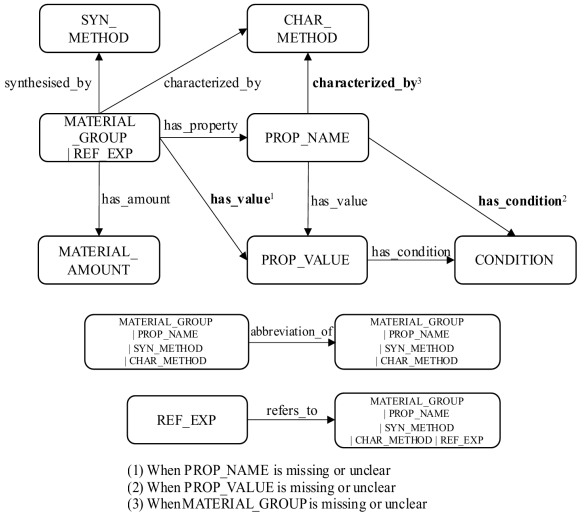


Figure 3: Illustration of the relationships between different polymer-related entities within PolyMinder.

```
{
  "text": "Sulfonated
     polyarylenethioethersulfone (SPTES
     ) copolymers with high ...",
  "entities": [
    [
      "T1",        #Entity ID
      "POLYMER",   #Entity Type
      [[0, 38]],   #Entity Position
      "Sulfonated
         polyarylenethioethersulfone"
                   #Entity Span
    ],
    [
      "T2",        #Entity ID
      "POLYMER",   #Entity Type
      [[41, 46]], #Entity Position
      "SPTES"      #Entity Span
    ],
    ...
  ],
  "relations": [
    [
      "R1",                #Relation ID
      "abbreviation_of", #Relation Type
      [["Arg1", "T2"], ["Arg2", "T1"]]
            #Start and End Entities
    ],
    ...
  ]
}
```

3.3 PolyMinder System

PolyMinder is a web-based application designed to facilitate the extraction, visualization, and annotation of polymer-related information from scientific documents. It integrates a Python-based backend with a JavaScript-based frontend, seamless interaction between the user and the system. This section details the core components of the system and describes the data flow, as illustrated in Figure 4.

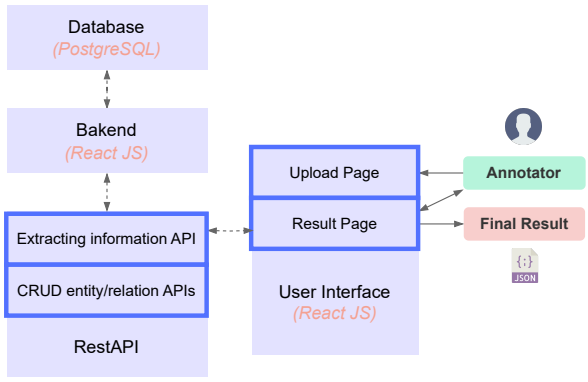


Figure 4: The architecture of the PolyMinder system, detailing its backend, RestAPI, and frontend.

3.3.1 Backend

The backend is implemented using the FastAPI framework[2] for its high performance and efficient development of RESTful APIs. These APIs handle communication with the frontend, supporting real-time data exchange and interaction. The backend's primary functions include document processing, data management, and facilitating user interactions.

When a PDF document is received, the backend employs PyMuPDF (McKie and Liu, 2020) to extract both text and its positional information within the document. PyMuPDF is particularly suited for this task because it supports the extraction of both text and bounding boxes, enabling accurate mapping of text to its location in the original PDF, which is crucial for visual annotations. The parsed text is then processed by pre-trained NER and RE models specifically tailored for polymer science, identifying and classifying relevant entities (e.g., polymer names, property names) and their relationships. For data management, the backend uses SQLAlchemy (Myers et al., 2015), an Object-Relational Mapping (ORM) tool that allows for flexible database selection, such as SQLite for lightweight applications or PostgreSQL for more demanding needs. The identified entities and relationships are stored in a structured, editable format, making it easy to retrieve and modify the data as needed. The backend also supports CRUD (Create, Read, Update, Delete) operations for entities, relationships, and PDF parsed text, enabling annotators to interact directly with the extracted data. Real-time updates are instantly reflected in the frontend through the REST APIs, ensuring a responsive and dynamic user experience that streamlines the annotation and correction process.

[2] https://fastapi.tiangolo.com/

3.3.2 Frontend

The frontend is developed using React[3] (JavaScript), along with HTML5 and CSS, to deliver an intuitive user interface. It allows users to upload polymer science PDFs for processing (Figure 5a) and visualizes extracted entities directly on the PDFs using overlays (Figure 5b), helping users see annotations in context. Relationships between entities are displayed in a Brat-like pop-up window (Stenetorp et al., 2012), offering clear insight into data connections (Figure 5c). To address potential errors from the PDF parser and NER/RE models, the frontend includes interactive tools allowing users to modify or correct parsed text and annotations, ensuring accuracy (Figure 5d). Upon completion, users can download the annotated documents for further analysis or use. Overall, the interface emphasizes ease of use and efficiency, streamlining annotators' workflows.

3.3.3 Data Flow

Figure 4 shows the typical workflow of the PolyMinder system. First, the user uploads a PDF document via the frontend interface. The backend processes the document using PyMuPDF to extract text and positional data. The extracted text is then sent to NER and RE models to identify relevant entities and their relations. The resulting data is stored in a database managed by SQLAlchemy, ensuring efficient retrieval and manipulation.

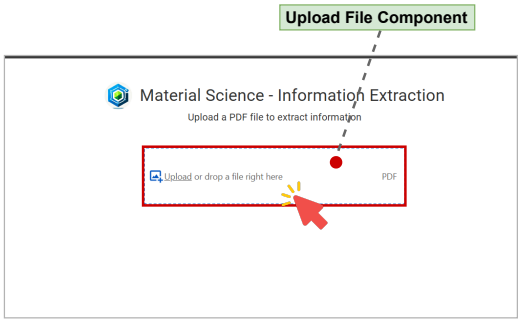
Next, the frontend accesses the extracted data via REST APIs and overlays the annotations on the original PDF, providing users with an intuitive visualization of the results. Users can review, refine, and edit the extracted entities, relationships, and parsed text using interactive tools. Any modifications made by the user are sent back to the backend through API calls, updating the database accordingly. If the parsed text data is edited, the system reprocesses the relevant components, generating and visualizing updated results. Once the user finalizes the annotations and is satisfied with the output, they can download the completed document, concluding the workflow.

4 Experiments on NER and RE models

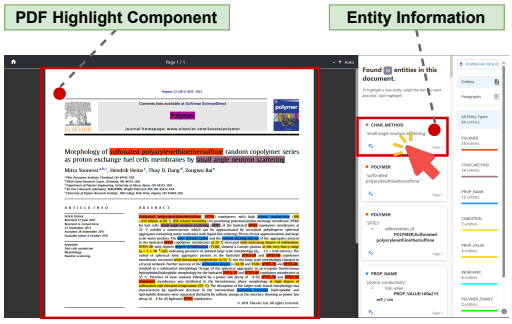
4.1 Dataset for NER and RE Tasks

Our NER and RE system consists of two modules, forming a pipeline to identify entity mentions and extract relations between them. Both models are

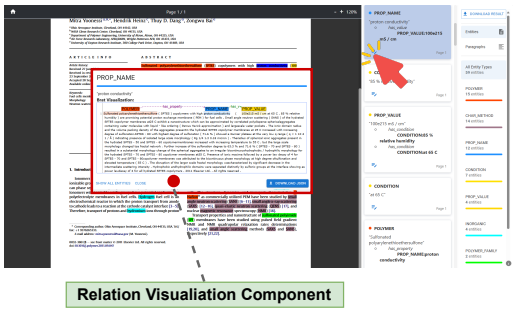
[3] https://react.dev/



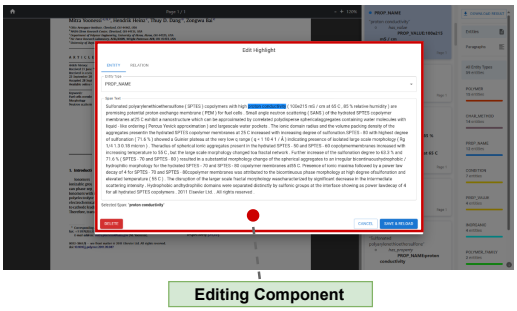
(a) Step 1: *On the upload page, users can drag and drop a document file to initiate the entity extraction process in the backend.*



(b) Step 2: *On the results page, users can click an entity in the sidebar to view its details and relationships.*



(c) Step 3: *On the results page, users can click the edit button next to an entity to open the editing panel for modification.*



(d) Step 4: *After editing, users can click the download button to save the final document.*

Figure 5: A step-by-step depiction of the typical user interaction flow within the PolyMinder interface, from document upload to entity editing and final result download. More guidelines are available on the demo website.

trained on the PolyNERE corpus (Phi et al., 2024) and are based on top-performing methods (Li et al., 2022; Zhou et al., 2021) in various configurations. For NER, most entities could be inferred from the context within the same sentence. Therefore, we focused on performing sentence-level NER to identify all possible entities in each abstract/paragraph. In contrast, extracting relationships between entities often requires cross-sentence reasoning, making it a paragraph or abstract-level task.

The PolyNERE corpus is split into 637 abstracts for training (85%), 38 for development (5%), and 75 for testing (10%). We report precision, recall, and F-score for both tasks. Models are developed and optimized using the training and development sets, with final evaluation on the test set.

4.2 Experimental Setup and Results

Named Entity Recognition. We conduct experiments using the W2NER model (Li et al., 2022), selected for its availability, efficiency, and ease of deployment. Furthermore, W2NER is particularly effective at identifying flat, overlapped, and discontinuous mentions, which are common in materials science texts where multiple entities are often discussed simultaneously. We utilize the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 1e-3. For the BERT-BiLSTM encoder layer, the model features a distribution embedding size of 20 and an LSTM hidden size of 1024. Dropout rates of 0.5 are applied for both embeddings and convolutions. Training runs for up to 50 epochs with a batch size of 12, and the best checkpoint based on the development set is saved after training. We also experimented with different encoders to enhance NER performance.

Table 1: Results for NER on test set

Method	Encoder	P	R	F1
W2NER (Li et al., 2022)	BERT-large	77.78	73.55	75.61
	SciBERT	74.89	75.67	75.28
	MatSciBERT	**78.05**	**76.53**	**77.28**

Table 1 shows the evaluation of NER on the test set. Table 1 demonstrates that MatSciBERT (Gupta et al., 2022) yields better performance compared to the use of SciBERT (Beltagy et al., 2019).

Relation Extraction. We define the RE task as a cross-sentence relation extraction problem and evaluate it with pre-defined gold entities. An entity may have multiple mentions in the abstract, and a relation between two entities (e1, e2) exists if

Table 2: Results for RE on test set given gold entities

Method	Encoder	P	R	F1
ALTOP (Zhou et al., 2021)	BERT-large	84.35	73.59	78.60
	SciBERT	83.59	81.60	82.58
	MatSciBERT	**83.99**	**82.49**	**83.23**

expressed by any pair of their mentions. During inference, the goal is to predict relations between all possible entity pairs. To achieve this, we adapted the ATLOP method (Zhou et al., 2021), which aggregates contextual information using Transformer attention and employs an adaptive threshold for different entity pairs. Since ATLOP operates at the document level (or, in this case, at the paragraph level), a postprocessing step is used to convert the results into binary relations between entity mentions. Our model is optimized using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 5e-5, a training batch size of 4, and a test batch size of 8, with a maximum of 30 epochs. We also experiment with different encoders.

As shown in Table 2, ATLOP achieves the highest F1 score of 83.23% using the MatSciBERT encoder. Overall, our system demonstrates strong performance and is well-suited for real-world use as an effective and practical RE system, addressing complex contexts in materials science papers, including flat, overlapping, and discontinuous entity mentions, as well as relations across sentences. Additionally, the NER and RE models integrated into PolyMinder are modular and replaceable, allowing customization with advanced tools or adaptation to other domains beyond polymer science.

4.3 Efficiency and Processing Time Analysis

The PolyNERE corpus consists of 750 abstracts, each containing an average of 25.24 entities and 15.29 relations. Based on an estimated annotation time of 15 seconds per item[4], factoring in the annotator's familiarity with materials science and the task's complexity, manually annotating a single abstract would take approximately 10 minutes. This estimate could increase due to challenges like lengthy paragraphs, complex entity relationships, and maintaining context across multiple sections.

To assess the efficiency improvements introduced by PolyMinder, we applied our system to the 75 abstracts in the test set. The total processing time was 6.45 minutes, averaging 5.16 seconds per abstract. This represents a significant reduction compared to the estimated 10 minutes required for manual annotation. Although verification and refinement time are not included in this figure, the high precision and recall of our NER and RE models (as demonstrated in Tables 1 and 2) suggest that the need for extensive post-processing is minimized. Even if an additional 3–4 minutes per abstract is allocated for review, the overall time remains well below the manual annotation time, presenting a considerable efficiency advantage for researchers handling large volumes of literature.

[4]Estimate provided by the PolyNERE corpus author.

5 Threats to Validity

While PolyMinder shows promise in supporting entity and relation annotation for polymer-related documents, several limitations may impact its generalizability and performance.

Dataset Size, Diversity, and Annotation Quality. A key limitation is the system's reliance on the PolyNERE corpus for training NER and RE models. Though tailored to the polymer domain, it contains only 750 abstracts, which may not represent the full diversity of polymer science literature. Additionally, data imbalance and incomplete annotations for some entities may lead to biased models that underperform on less frequent or poorly labeled entities. Future work will focus on expanding the dataset with more diverse documents and improving annotation quality to boost robustness.

PDF Extraction Inconsistencies. Variations in PDF formatting, such as complex layouts, figures, and tables, create challenges. These inconsistencies can result in extraction errors, causing missed or incorrect entity annotations. Future work will investigate advanced extraction techniques to better handle diverse PDF structures.

6 Conclusion

In this paper, we introduced PolyMinder, a specialized support system that streamlines entity extraction and relation annotation in polymer science documents by leveraging advanced NER and RE models tailored specifically to the polymer domain. Our system automates the extraction of key polymer-related entities and their relationships, providing an intuitive web interface for users to efficiently browse, verify, and refine the information. Experimental results demonstrate high performance on the PolyNERE corpus, highlighting efficiency gains over manual annotation processes.

References

Mariam Al Ali AlMaadeed, Deepalekshmi Ponnamma, and Ali Alaa El-Samak. 2020. Polymers to improve the world and lifestyle: physical, mechanical, and chemical needs. In *Polymer Science and Innovative Applications*, pages 1–19. Elsevier.

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.

Ekaterina Borisova, Raia Abu Ahmad, Leyla Garcia-Castro, Ricardo Usbeck, and Georg Rehm. 2024. Surveying the FAIRness of annotation tools: Difficult to find, difficult to reuse. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 29–45, St. Julians, Malta. Association for Computational Linguistics.

Jerry Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2024. POLYIE: A dataset of information extraction from polymer material scientific literature. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2370–2385, Mexico City, Mexico. Association for Computational Linguistics.

Rifkat Davronov and Fatima Adilova. 2023. Bert-based drug structure presentation: A comparison of tokenizers. In *AIP Conference Proceedings*, volume 2781. AIP Publishing.

Danielle E Fagnani, Coralie Jehanno, Haritz Sardon, and Anne J McNeil. 2022. Sustainable green polymerizations and end-of-life treatment of polymers.

Tanishq Gupta, Mohd Zaki, NM Anoop Krishnan, and Mausam. 2022. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10965–10973.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

M. McKie and R. Liu. 2020. Pymupdf - python binding for mupdf. `https://pypi.org/project/PyMuPDF/`. Accessed: 2024-09-12.

Amar K Mohanty, Feng Wu, Rosica Mincheva, Minna Hakkarainen, Jean-Marie Raquez, Deborah F Mielewski, Ramani Narayan, Anil N Netravali, and Manjusri Misra. 2022. Sustainable polymers. *Nature Reviews Methods Primers*, 2(1):46.

Jason Myers, Rick Copeland, and Richard D Copeland. 2015. *Essential SQLAlchemy*. " O'Reilly Media, Inc.".

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Hiroyuki Oka, Atsushi Yoshizawa, Hiroyuki Shindo, Yuji Matsumoto, and Masashi Ishii. 2021. Machine extraction of polymer data from tables using xml versions of scientific articles. *Science and Technology of Advanced Materials: Methods*, 1(1):12–23.

Obinna Okolie, Anuj Kumar, Christine Edwards, Linda A Lawton, Adekunle Oke, Seonaidh McDonald, Vijay Kumar Thakur, and James Njuguna. 2023. Bio-based sustainable polymers and materials: From processing to biodegradation. *Journal of Composites Science*, 7(6):213.

Van-Thuy Phi, Hiroki Teranishi, Yuji Matsumoto, Hiroyuki Oka, and Masashi Ishii. 2024. PolyNERE: A novel ontology and corpus for named entity recognition and relation extraction in polymer science domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12856–12866, Torino, Italia. ELRA and ICCL.

Shubham Sharma, P Sudhakara, Abdoulhdi A Borhana Omran, Jujhar Singh, and RA Ilyas. 2021. Recent trends and developments in conducting polymer nanocomposites for multifunctional applications. *Polymers*, 13(17):2898.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52.

Hiroyuki Shindo, Yohei Munesada, and Yuji Matsumoto. 2018. PDFAnno: a web-based linguistic annotation tool for PDF documents. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mohammad Golam Sohrab, Khoa Duong, Makoto Miwa, Goran Topić, Ikeda Masami, and Takamura Hiroya. 2020. BENNERD: A neural named entity

linking system for COVID-19. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 182–188, Online. Association for Computational Linguistics.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Matthew C Swain and Jacqueline M Cole. 2016. Chemdataextractor: a toolkit for automated extraction of chemical information from the scientific literature. *Journal of chemical information and modeling*, 56(10):1894–1904.

Neelanshi Wadhwa, S Sarath, Sapan Shah, Sreedhar Reddy, Pritwish Mitra, Deepak Jain, and Beena Rai. 2021. Device fabrication knowledge extraction from materials science literature. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15416–15423.

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of chemical information and modeling*, 59(9):3692–3702.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19):9691.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.