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

Dinh-Truong Do[1*] **Hoang-An Trieu**[1*] **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, synthesis methods, and material properties. 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 automated support system designed to assist polymer scientists in extracting and annotating polymer-related entities and their relationships from scientific documents. The system utilizes state-of-the-art 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 visualize, verify, and refine the extracted information before finalizing the annotations. 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). Polymers possess remarkable versatility, offering a wide range of properties that make them indispensable in the development of essential products such as plastics, rubbers, adhesives, and nanomaterials (Mohanty et al., 2022; AlMaadeed et al., 2020). As research in this area continues to grow, the volume of scientific literature containing

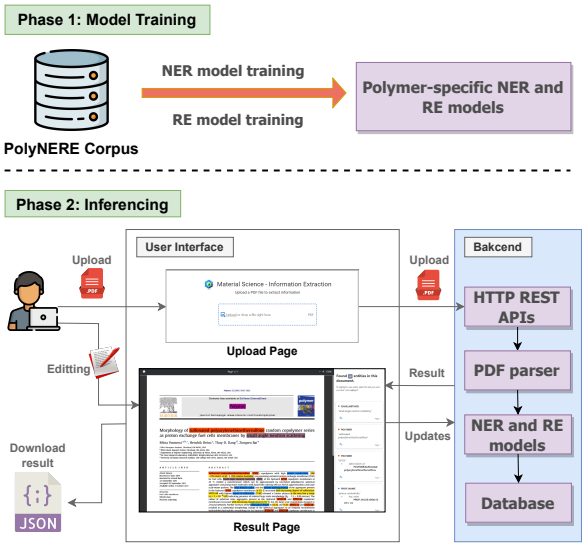


Figure 1: A comprehensive overview of the PolyMinder system, showcasing its workflow from model training to entity extraction and user interaction via the web-based interface.

critical insights into polymer structures, properties, and applications has expanded rapidly (Phi et al., 2024). Effectively organizing and extracting valuable information from this vast body of work is vital for advancing both research and industrial applications. However, manually extracting and annotating domain-specific entities—such as polymer names, synthesis methods, and properties—presents a significant challenge (Fagnani et al., 2022). The complexity of scientific language, coupled with the specialized expertise needed for accurate interpretation, makes manual annotation time-consuming and prone to errors.

Automated named entity recognition systems have seen significant success in the materials domain, and polymers are no exception. Previous works have developed systems that extract polymer-related entities using advanced deep neural networks, yielding promising results (Oka et al., 2021; Phi et al., 2024; Cheung et al., 2024). However,

[*]Equal contribution

these systems typically output results in text-based formats like JSON, without providing intuitive visualizations for users or annotators to easily review the extracted information. Moreover, while these systems rely on neural network models to extract material entities, the models are not perfect and can produce errors. Current systems also lack features that allow annotators to refine or correct the extracted data efficiently, requiring 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 texts—such as polymer names, properties, synthesis methods, and experimental conditions—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 state-of-the-art 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 as follows:

- **Creation of PolyMinder**: An automated support system that extracts and visualizes key polymer-related information from scientific texts, allowing annotators to review and refine the extracted data.
- **Polymer-Specific Entity and Relation Extraction Models**: Custom models tailored to polymer science, utilizing SOTA techniques.
- **Public Release of Source Code**[1] : We make PolyMinder's source code publicly available to support further research and development.

The rest of this paper is organized as follows: Section 2 reviews related work on automated entity extraction in materials science and polymers. Section 3 outlines PolyMinder's methodology, including model training and system architecture. Section 4 presents the experimental setup and results, while Section 5 discusses limitations and future research directions. .

2 Related Works

Automated information extraction from scientific literature has significantly advanced in domains such as biomedical science, chemistry, and materials science (Gupta et al., 2022; Olivetti et al., 2020; Nasar et al., 2018; Krallinger et al., 2017; Rocktäschel et al., 2012). These advancements have led to the development of specialized systems capable of efficiently extracting valuable information from unstructured text, thereby accelerating research and innovation. In polymer science, Oka et al. (2021) developed a system for extracting polymer data from tables in scientific articles. By integrating a deep neural network for polymer name recognition with rule-based algorithms for property identification, their approach enhances the accuracy and efficiency of data retrieval in polymer research. Similarly, Swain and Cole (2016) introduced ChemDataExtractor, an automated tool for extracting chemical data such as molecules, reactions, and material properties from unstructured text. The system combines rule-based methods with supervised learning to process large volumes of chemical literature, supporting data-driven discovery in chemistry. To address the challenge of processing the extensive materials science literature, Weston et al. (2019) developed MatScholar. This tool employs named entity recognition to extract essential information from articles, focusing on inorganic materials, sample descriptions, phase labels, material properties, and synthesis or characterization methods. MatScholar aids researchers in quickly identifying relevant information, streamlining the research process in materials science. In the biomedical domain, the BENNERD system (Sohrab et al., 2020) specializes in extracting entities related to COVID-19. Leveraging the CORD-NER dataset (Wang et al., 2020) for pre-training, BENNERD provides a real-time entity annotation and linking platform, which is critical for rapid information dissemination during global health crises. Furthermore, Wadhwa et al. (2021) proposed a system for extracting entities and relations within materials science literature, focusing on device fabrication

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

knowledge. Their system utilizes state-of-the-art models for entity and relation extraction to identify key entities such as operations, materials, and methods, uncovering relationships between these elements. This approach enables the mapping of comprehensive fabrication processes, contributing significantly to advancements in materials science research.

Despite these advancements, existing systems often lack intuitive interfaces for data refinement. Many tools output extracted information in formats like JSON without user-friendly visualizations, making it difficult for researchers to review and interact with the data. Neural network models can produce errors, yet current platforms do not allow users to efficiently correct these within the system. Therefore, there is a need for more adaptable information extraction frameworks that accurately extract domain-specific entities and relations while providing interactive interfaces for visualization and refinement, enhancing data annotation efficiency in polymer science.

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 (NER) and Relation Extraction (RE) models using state-of-the-art 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.

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 is optimized for fine-grained polymer-related NER. W2NER utilizes a word-to-span alignment approach, which allows it to handle overlapping and nested entities more 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 16 entity types. For example, entity POLYMER related to material entities that are polymers, for example, *"Sulfonated poly(phthalazinone ether ketone nitrile)"*. All entity types used in our system are summarized in Figure 2.

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

Figure 2: Ontology of material entities in PolyMinder, illustrating various material types. Monomers are categorized under organic, while polymers can be organic or inorganic.

After identifying the entities, the next step is to establish relationships between them. In polymers science, these relationships capture the fundamental connections between the polymer-related entities, for example the entity POLYMER has relation *characterized_by* to the CHAR_METHOD entity. The details of all entities and their relationships in our system are illustrated in Figure 3. TO extract these relations, we approach RE task at the paragraph level, focusing on predicting relationships between entity pairs across the text. For this, we adapt the ATLOP model (Zhou et al., 2021), which is tailored for document-level or paragraph-level RE. By leveraging transformer-based attention mechanisms, ATLOP captures complex, cross-sentence relationships.

Once trained, these NER and RE models become integral components of the backend infrastructure within the PolyMinder system, producing output in JSON format. An example of this output is as follows:


```
{
  "text": "Sulfonated
      polyarylenethioethersulfone (SPTES
      ) copolymers with high ...",
  "entities": [
```

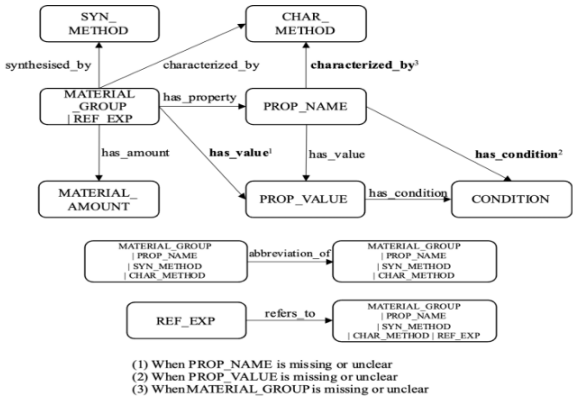


Figure 3: Illustration of the relationships between different polymer-related entities within the PolyMinder system.

```
    [
      "T1",        #Entity ID
      "POLYMER",   #Entity Type
      [[0, 38]],   #Entity Position
      "Sulfonated
          polyarylenethioethersulfone"
                   #Entity Span
    ],
    [
      "T2",
      "POLYMER",
      [[41, 46]],
      "SPTES"
    ],
    ...
  ],
  "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-powered backend with a JavaScript-based frontend, ensuring seamless interaction between data processing and user interface components. This section details the core components of the system and describes the data flow, as illustrated in Figure 4.

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.

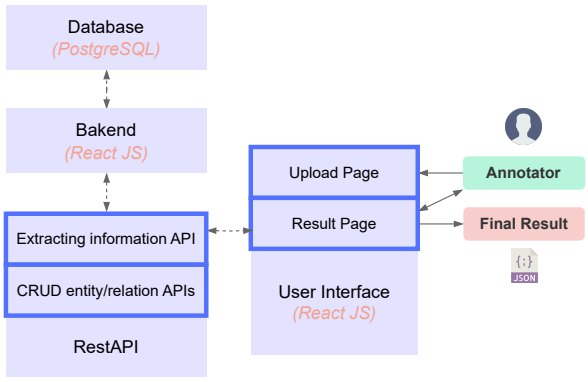


Figure 4: The architecture of the PolyMinder system, detailing its backend infrastructure, API communication, and frontend interaction.

Upon receiving a PDF document, the backend utilizes PyMuPDF (McKie and Liu, 2020) to extract textual content along with its positional information within the document. This enables accurate mapping of text to its location in the original PDF, which is crucial for visual annotations. The extracted text is then processed by pre-trained NER and RE models that are specifically tailored for polymer science. These models identify and classify relevant entities (e.g., polymer names, properties) and their interrelationships within the text.

For data management, the backend employs SQLAlchemy (Myers et al., 2015), an Object-Relational Mapping (ORM) tool that offers flexibility in database selection, such as SQLite[3] for lightweight applications or PostgreSQL[4] for more robust requirements. Extracted entities and relationships are stored in a structured, editable format within the database, facilitating efficient retrieval and modification.

The backend also supports CRUD (Create, Read, Update, Delete) operations for entities, relationships, and paragraph-level text. This allows annotators to interact directly with the extracted data, making real-time updates that are immediately reflected in the frontend through the REST APIs. This design ensures a responsive and dynamic user experience, enabling annotators to efficiently refine and correct the extracted information.

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

[3] https://www.sqlite.org/

[4] https://www.postgresql.org/

3.3.2 Frontend

The frontend is developed using React[5] (JavaScript), along with HTML5 and CSS, to deliver an intuitive and responsive user interface. The frontend allows users to upload polymer science documents in PDF format for processing, and provides visualization by displaying extracted entities directly on the PDF using overlays, helping users see annotations in context. Relationships between entities are also displayed, enabling a clear understanding of the data connections. Users can interact with the system by modifying or correcting annotations through intuitive editing tools, ensuring accuracy. After the annotation process is complete, users can download the finalized, annotated documents for further use or analysis.

The frontend's design emphasizes ease of use and efficiency, aiming to streamline the annotators' workflow (Figure 5). By offering immediate visual feedback and interactive editing features, the system helps users quickly identify and correct any inaccuracies in the automated extraction, thus enhancing the overall quality of the annotations.

3.3.3 Data Flow

Figure 4 illustrates the typical workflow of the PolyMinder system. Initially, the user uploads a PDF document through the frontend interface. The backend then processes the document using PyMuPDF to extract text and positional data. Once extracted, the text is sent to Named Entity Recognition (NER) and Relationship Extraction (RE) models to identify and classify relevant entities and their relationships. The resulting data is stored in a database managed by SQLAlchemy, ensuring efficient data 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 and refine the extracted entities and relationships using interactive editing tools, and any changes made are sent back to the backend through API calls, updating the database. Once the user is satisfied with the annotations, they can download the finalized document, completing the workflow.

This process promotes efficiency by allowing real-time corrections and leveraging a user-friendly interface, which significantly enhances both the speed and accuracy of data annotation in polymer science.

[5] https://react.dev/

4 Entity and Relation Extraction Experiments

4.1 Experimental Settings

4.2 Results

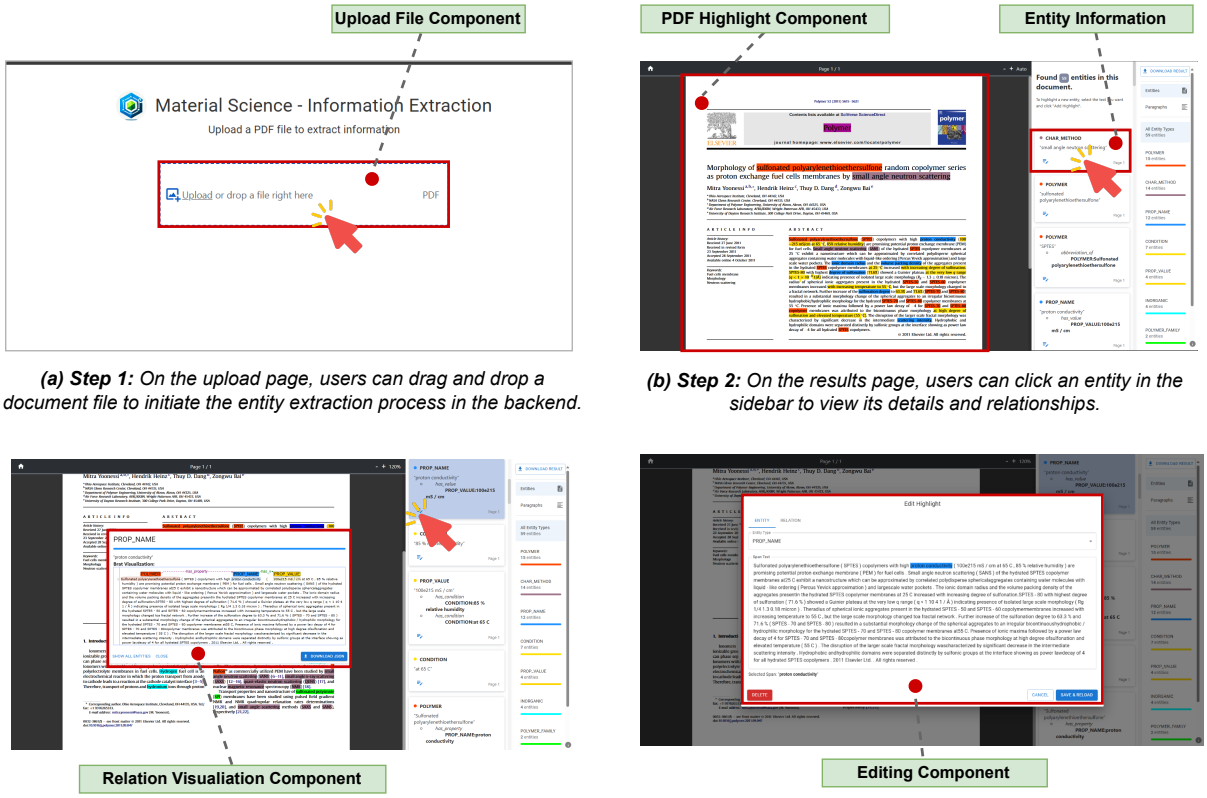
5 Threats to Validity

While the PolyMinder system shows promise in supporting entity and relation annotation for polymer-related documents, several limitations could impact its generalizability and performance. **Domain-specific dataset limitations** could hinder the system's ability to generalize. PolyNERE is tailored to polymer terminology, which may not fully cover emerging subdomains or new terms in material science. This may reduce accuracy when processing documents with novel vocabulary. *Future work* will focus on expanding the dataset to cover more subdomains and continuously updating the terminology.

PDF extraction inconsistencies present a challenge due to the variability in PDF formatting, such as complex layouts, figures, or tables. These inconsistencies can cause extraction errors, leading to missed or incorrectly annotated entities. *Future work* will explore more advanced extraction techniques that handle diverse PDF structures better.

6 Conclusion

In this paper, we presented PolyMinder, an automated support system for polymer-related entity and relation extraction from scientific literature. By leveraging state-of-the-art models and a domain-specific dataset, PolyMinder addresses the unique challenges associated with processing polymer science documents. The system not only automates the extraction process but also empowers domain experts to verify and refine the outputs through an intuitive web interface, ensuring accuracy and reliability in the final annotations. Our experimental results demonstrate the system's potential to significantly reduce manual annotation effort while improving the overall consistency and quality of extracted data. Future work will focus on expanding the dataset to cover emerging subdomains and improving the system's handling of complex document structures to further enhance its applicability. The open-source release of the system's code will



(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.

support the community in advancing automated information extraction in polymer science and other specialized fields.

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.

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.

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.

Martin Krallinger, Obdulia Rabal, Analia Lourenco, Julen Oyarzabal, and Alfonso Valencia. 2017. Information retrieval and text mining technologies for chemistry. *Chemical reviews*, 117(12):7673–7761.

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.

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.".

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2018. Information extraction from scien-

tific articles: a survey. *Scientometrics*, 117(3):1931–1990.

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.

Elsa A Olivetti, Jacqueline M Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M Hiszpanski. 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4).

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.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. Chemspot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

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.

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.

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.

Xuan Wang, Xiangchen Song, Yingjun Guan, Bangzheng Li, and Jiawei Han. 2020. Comprehensive named entity recognition on cord-19 with distant or weak supervision. *ArXiv*, abs/2003.12218.

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.

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.