A Practical and Customizable System for Named Entity Recognition and Relation Extraction in Materials Science Publications

Van-Thuy Phi and Yuji Matsumoto

Center for Advanced Intelligence Project, RIKEN
{thuy.phi, yuji.matsumoto}@riken.jp

Abstract. Materials science encompasses a wide variety of entities ranging from high-level categories like organic and inorganic to lower-level subcategories such as polymers. With the growing number of publications in this field, there is a rising need for automated methods to extract information on findings from relevant publications. In this paper, we introduce MatSciNERE, a comprehensive corpus curated for a wide range of material entities and their relationships within the materials science domain. Utilizing a material-relevant ontology, we have developed a practical Named Entity Recognition (NER) and Relation Extraction (RE) pipeline. This system is specifically designed to identify fundamental concepts and relationships in material science, with the flexibility to expand its functionality for future research on specific entities and relations of interest. Emphasizing practical usage, our system is adept at handling the complex and varied contexts present in scientific material papers. By conducting evaluation and result analysis at both abstract and paragraph levels, our system demonstrates readiness for real-world scenarios as an effective and practical RE system. We plan to provide a corpus of 1,000 annotated abstracts and paragraphs to support research in this field.

Keywords: Named Entity Recognition, Relation Extraction, Materials Science Corpus, Polymer Science Corpus.

1 Introduction

The field of materials science covers a diverse range of entities, spanning broad classifications like organic, inorganic, and composite, as well as more specific subcategories like polymers. Given the increasing volume of publications in this field, there is a growing demand for automated methods to assist researchers in extracting valuable information from scientific publications. These findings vary across papers and depend on the specific requirements of researchers, including details like property names and corresponding values of materials, among other relevant information.

Despite the growing demand for automated methods to efficiently extract information from scientific papers, the availability of resources and practical information extraction systems that can provide accurate predictions for the broader

materials science domain, while also being adaptable to specific subdomains, is severely limited. Existing resources typically provide annotations exclusively for entities (Weston et al., 2019; Yamaguchi et al., 2020; O'Gorman et al., 2021; Shetty et al., 2023), while a limited number include annotations for both entities and relations, either within the general materials science domain (Mysore et al., 2019) or specific subdomains (e.g., polycrystalline materials, Yang et al., 2022). Despite the prevalence of fundamental entity categories like organic and inorganic, there are no annotations for both of these entity types in the text to the best of our knowledge. This absence of resources and practical information extraction systems significantly hinders the speed at which knowledge can be acquired from materials science literature.

To address this gap, we present MatSciNERE, a comprehensive corpus curated for a wide range of material entities and their relationships within the materials science domain. The MatSciNERE corpus is constructed based on an earlier version of our corpus in prior research, where we explored the Named Entity Recognition (NER) and Relation Extraction (RE) tasks in the polymer science domain. The annotation assumption previously used has been adapted to create a more comprehensive resource, aiding in the improvement of the corpus from polymer science to material science. With a high number of entity and relation annotations, the research and development of information extraction methods from materials science publications could potentially benefit from our corpus.

Based on a material-centric ontology, we also develop a practical NER and RE system specifically designed to identify fundamental concepts and relationships in material science (such as between broad entities like organic or inorganic), with the flexibility to expand its functionality for future research on specific entities and relations of interest (such as between relevant entities in the polymer science domain).

Our system consists of two modules functioning as a pipeline for recognizing entity mentions and extracting relation pairs between them. Both the NER and RE modules are experimentally developed based on top-performing methods under various settings. Our system effectively addresses complex contexts in materials science papers, handling flat, overlapped, and discontinuous entity mentions, along with relations across sentences. Through extensive evaluation on numerous abstracts and paragraphs, it demonstrates its effectiveness for real-world scenarios, aiming to consistently produce accurate predictions as a practical end-to-end RE system.

The main contributions of this paper are as follows:

- We introduce MatSciNERE, a large and high-quality corpus for the Materials Science domain containing Named Entities and Relations. This corpus consists of fourteen entity types and eight relation types, with high annotation coverage. It is constructed based on an earlier version of our corpus in prior research focusing on the polymer science domain.
- We have developed a pipeline system comprising two modules, NER and RE. Both the corpus and the system prioritize practical usage, designed to handle complex expressions and diverse contexts, particularly those related to overlapped and discontinuous entity mentions. The system extracts target relations beyond the sentence boundary. Moreover, the importance of abbreviations and referring

expressions is essential in interpreting the behavior of our Relation Extraction (RE) model. We explicitly incorporate these elements into our relation schema, eliminating the need for external tools like abbreviation detectors or coreference resolution tools.

- Our system employs top-performing NER and RE methods, with various evaluations emphasizing its strong performance in extracting target entities and relations. We also assess the system by extracting information from 250 paragraphs in different materials science papers and analyzing its predictions. We plan to release the MatSciNERE corpus, which includes 1,000 annotated abstracts and paragraphs, and the trained models to support research in the material science as well as polymer science field.

2 Materials Science Corpus Construction

2.1 Our Prior Polymer Corpus

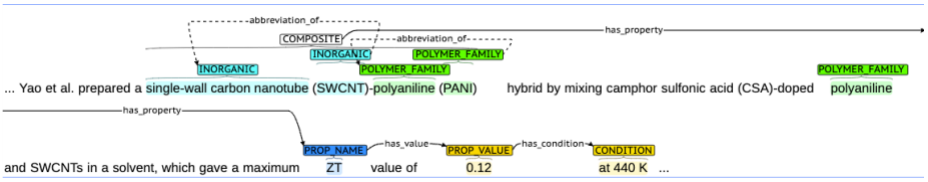
Prior to this work, we have constructed a polymer corpus for abstract-level NER and RE from plain text, which captures essential information about polymer-related entities and their relations frequently found in polymer and materials abstracts. Our PolyNERE corpus consists of 750 polymer paper abstracts with text sources closely matching the PolymerAbstracts corpus (Shetty et al. 2023).

Our PolyNERE corpus contains fourteen entity types and eight relation types, and was designed as a resource for developing an NER and RE system, focusing on polymer and their property information.

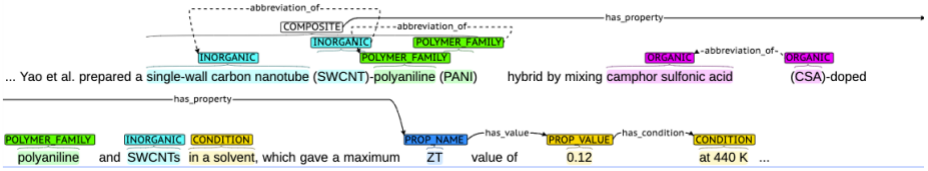
2.2 Construction of Materials Science Corpus

Based on the PolyNERE corpus, we obtained good performance on identifying polymer names, and their associated property names and values. For instance, F1 scores for recognizing these entities are all higher than 80% in our evaluation. However, a major issue to prevent deploying this system is related to the low F1 scores for some entities other than the above types, which can be caused by the data imbalanced issue, and the incomplete annotations for some of the entity types.

We believe that the underlying issue lies in the annotation assumption we used to annotate the corpus. In the PolyNERE corpus, only entities and relations relevant to our target domain were annotated. As a result, in Figure 1a, even though the second '*SWCNTs*' mention is an inorganic material, it was not provided a label. We argue that it greatly affects the overall performance of NER and RE systems trained on those annotations, and limits the practical usage of these systems. Therefore, in this work, we made an important change in the annotation assumption by considering all entity mentions the text, aiming at a practical RE system for both general material science and its subdomains like polymer science. Figure 1b illustrates our new annotation assumption applied to our newly developed MatSciNERE corpus.



(a) Annotations specific to the field of polymer science



(b) Annotations applicable to the broader domain of materials science

Fig. 1. Annotation assumption (a) In PolyNERE corpus: only entities and relations relevant to the target domain were annotated; the first '*SWCNTs*' was labeled as it is part of a composite with another polymer class. (b) In MatSciNERE corpus: other labeled entities include '*camphor sulfonic acid*', '*CSA*', the second '*SWCNTs*' mention, '*in a solvent*', and another '*abbreviation_of*' relation.

We use the same ontology of all entities as for PolyNERE, as shown in Figure 2, where entity types are highlighted in bold text. Wherever hierarchies of concepts are present (e.g., ORGANIC→MONOMER), it is desirable to annotate the entity with the more specific type (i.e., MONOMER) unless it is unclear from the context.

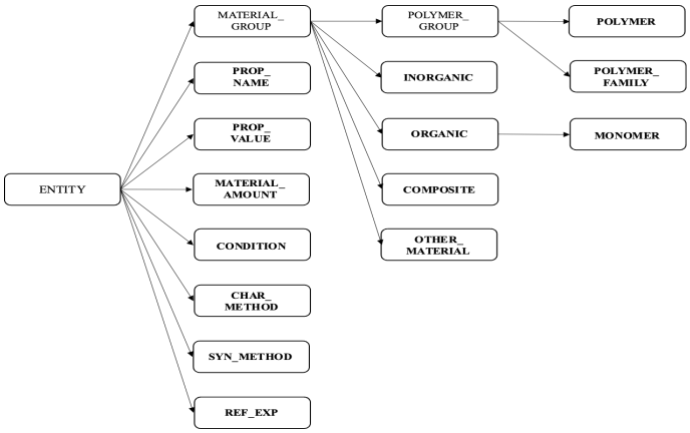


Fig. 2. Ontology of material entities separates types by their definitions; monomers are organic; polymers can be inorganic or organic.

In our ontology, MATERIAL_GROUP is defined as the group that contains POLYMER, POLYMER_FAMILY, MONOMER, ORGANIC, INORGANIC, COMPOSITE, and OTHER_MATERIAL. Figure 3 shows the relation schema that we use for both of our corpora. Our ontology of entities and relation schema can serve as

a foundation for extending the development of general-purpose information extraction systems in other subdomains, e.g., for metallic materials, bio-based materials, etc.

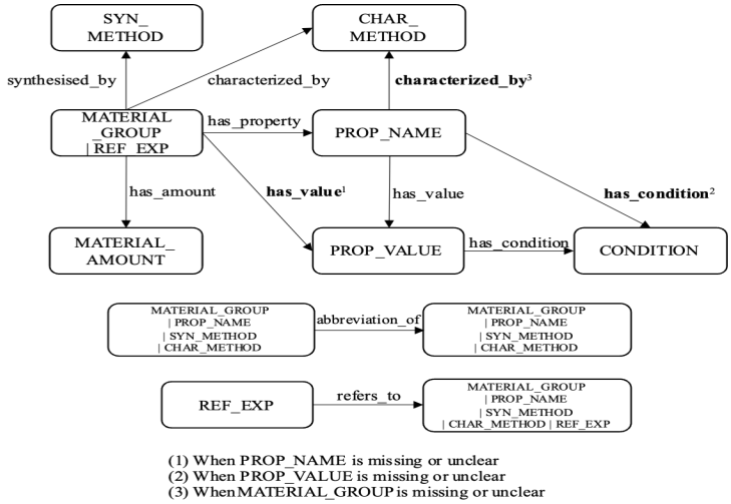


Fig. 3. Illustration of material entity relationships.

We use eight relation types as follows: *has_property*, *has_value*, *has_amount*, *has_condition*, *synthesised_by*, *characterized_by*, *abbreviation_of* and *refers_to*. The initial six relations in our corpus capture fundamental connections between the entities, typically within the sentence level. In contrast, the *abbreviation_of* and *refers_to* relations provide reasoning abilities to the RE task by offering supporting evidence across multiple sentences. Our relations are designed so that they can capture crucial and fine-grained information about material entities.

2.3 Annotation Process

The annotation was carried out using the BRAT tool (Stenetorp et al., 2012) which allows for annotating flat, overlapped, and discontinuous mentions. For instance, to annotate for discontinuous entities in the following sentence "*Photoluminescence maxima of P1, P2 and P3 films are 564, 559 and 558 nm, respectively*.", three property values are annotated: "*564 nm*", "*559 nm*" and "*558 nm*".

Our MatSciNERE corpus consists of 750 polymer relevant abstracts, each accompanied by its raw text and DOI information. A single annotator labeled all the entities and relations to ensure maximal coherence in the entity-relation schema. This approach was also used in annotating widely used datasets like Matscholar (Weston et al., 2019).

Our annotation process, consisting of three main rounds, is described as follows:
1st round: We annotate entities such as POLYMER, POLYMER_FAMILY, MONOMER, ORGANIC, INORGANIC, and MATERIAL_AMOUNT by referencing the PolymerAbstracts corpus (Shetty et al., 2023). Our aim is to enhance the accuracy

of annotations by adding, removing, or modifying inconsistent mentions. We also annotate PROP_NAME and PROP_VALUE according to our definitions and requirements for more precise mentions, including phrases like "*around*", "*higher than*", etc. Our focus extends to more specific property names and values.

Furthermore, we carry out annotations for the six entity types (CONDITION, SYN_METHOD, CHAR_METHOD, COMPOSITE, OTHER_MATERIAL and REF_EXP) incorporated into our ontology.

Under our new annotation assumption, as depicted in Figure 1b, we have expanded the scope of material entities. Annotation coverage for all entity types is as follows:

- **PROP_NAME**: All specific property names
- **PROP_VALUE**: All corresponding absolute property values
- **MATERIAL_AMOUNT**: To the maximum extent feasible
- **COMPOSITE**: To the maximum extent feasible
- **OTHER_MATERIAL**: Essential mentions related to target relations
- **CONDITION**: To the maximum extent feasible for train set; all mentions for dev/test set
- **REF_EXP**: Essential referring expressions required for reasoning target relations
- Other entity types (**POLYMER**, **POLYME_FAMILY**, **MONOMER**, **ORGANIC**, **INORGANIC**, **SYN_METHOD**, and **CHAR_METHOD**): All mentions

2nd round: Our emphasis is on annotating eight types of relations, along with three special relations designed to handle complex and varied contexts (e.g., coordination structures) used to describe entities relevant to polymers and their relationships.

3rd round: We conduct a re-check to ensure data consistency. This includes addressing issues such as overly generic entity mentions and the removal of certain relation pairs which involve those entities.

Also, in each round, we refine annotation guidelines and ensure consistent annotations through ongoing discussions between the annotator and a polymer expert. The annotator seeks guidance from the polymer expert when necessary, and revisions are primarily finalized after the third round.

2.4 Corpus Statistics

Our MatSciNERE corpus consists of a total of 750 abstracts, divided into three sets: 637 for training, 38 for development, and 75 for testing. Table 1 displays the statistics for our corpus, presenting details about the annotation type, and the number of annotations across various categories within MatSciNERE. Overall, our MatSciNERE corpus provides a rich source of information for training and evaluating models in the field of polymer science, particularly for tasks related to NER and RE.

Table 1. General statistics for our corpus.

#tokens/sentence	11.87
#entities/abstract	29.73
#relations/abstract	15.91
Overlapped entities	3,490 mentions (15.65%)
Discontinuous entities	281 mentions (1.26%)
ENTITY (14)	Total: 22,296 mentions
POLYMER	4,053 (582/750 abstracts)
POLYMER_FAMILY	1,159 (315)
PROP_NAME	3,882 (717)
PROP_VALUE	1,829 (587)
MONOMER	1,600 (320)
ORGANIC	1,855 (435)
INORGANIC	1,939 (393)
MATERIAL_AMOUNT	539 (267)
COMPOSITE	398 (172)
OTHER_MATERIAL	258 (120)
CONDITION	1,376 (552)
SYN_METHOD	381 (231)
CHAR_METHOD	1,752 (435)
REF_EXP	1,275 (460)
RELATION (8 +3 special)	Total: 11,935 pairs
has_property	3,502 (661/750 abstracts)
has_value	1,903 (582)
has_amount	424 (225)
has_condition	1,104 (406)
synthesised_by	282 (193)
characterized_by	1,347 (391)
abbreviation_of	2,033 (627)
refers_to	1,340 (459)

MatSciNERE contains 22,296 entity mentions, which is 2.05 times higher than the number found in PolymerAbstracts corpus. There are 3,490 overlapped entity mentions, constituting 15.65% of all entity mentions. The number of discontinuous entity mentions is 281, representing 1.26% of all mentions. While the proportion of discontinuous mentions is relatively low, it is worth noting that crucial entities associated with property information such as PROP_VALUE and PROP_NAME still include such mentions. Moreover, the total count of relation pairs is 11,935.

To assess the quality of the corpus, we randomly selected 10 abstracts from the test set, in which only the annotator was involved in all annotation rounds. We then compare the annotator's annotations with the corresponding annotations provided by a polymer expert. The true positives (tp), false positives (fp), and false negatives (fn) were determined to be 287, 8, and 85, respectively. Using these annotation statistics, we computed the precision, recall, and F1 scores, resulting in the following scores: P=97.29%, R=77.15%, and F1=86.06%. We achieved a Cohen's Kappa coefficient of 0.819.

3 Our NER and RE System

Our NER and RE system consists of two modules functioning as a pipeline for identifying entity mentions and extracting relation pairs between them. Both the NER and RE modules are experimentally developed based on top-performing methods under various settings.

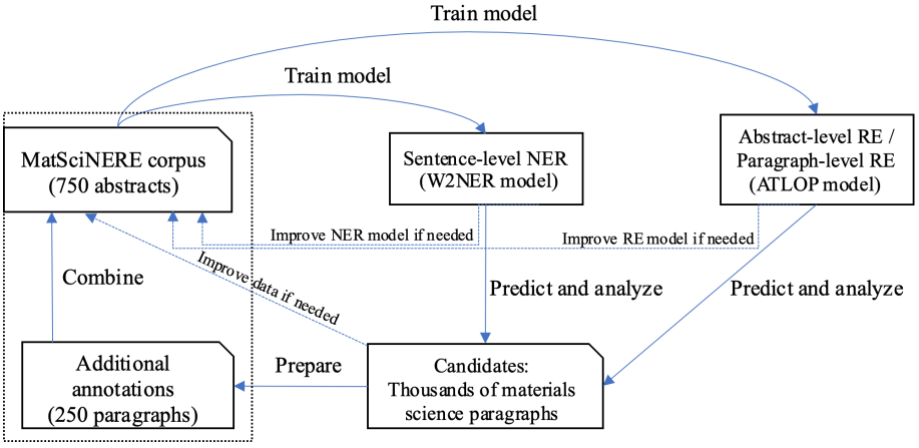


Fig. 4. Our practical NER and RE system.

Figure 4 depicts our system, where both the NER and RE models are trained using the MatSciNERE corpus. These models have been gradually enhanced through our analysis of results from a high number of candidates of materials science paragraphs. If necessary, the MatSciNERE corpus will be expanded to improve the quality and quantity of annotations.

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.

On the other hand, the identification of relationships between entities requires cross-sentence reasoning, which is equivalent to abstract-level or paragraph-level RE.

The MatSciNERE corpus is divided into three sets: 637 abstracts for training, 38 for development, and 75 for testing. Standard precision, recall and F-score metrics are reported for both NER and RE. Initially, the training and development sets are used for model development and parameter optimization, followed by the evaluation of a trained model on the test set.

3.1 NER Performance

We conduct experiments using the following methods: Span-based (Li et al., 2021), Transition-based (Dai et al., 2020), MaxClique (Wang et al., 2021), BARTNER (Yan et al., 2021) and W2NER (Li et al., 2022). These models can handle all flat, overlapped, and discontinuous mentions.

To ensure a fair comparison, we employ the BERT-large encoder (Devlin et al., 2019) for all experiments and only BART-large (Lewis et al., 2020) for the BARTNER model. Our default optimizer is Adam (Kingma and Ba, 2015), supplemented with a linear warmup and linear decay learning rate schedule. Our experiments are conducted using a batch size of *8* and run for a total of *30* training epochs. We follow similar settings for other hyperparameters, such as the learning rate, etc. in each baseline.

Table 2 shows the evaluation of NER on the test set. The W2NER model achieves the best performance on MatSciNERE, achieving the highest F1 score (75.61%) in comparison to other approaches. It outperforms the previous best F1 score, represented by the MaxClique model, by a margin of 1.1%. The BARTNER and Transition-based models achieved an F1 of 74.02% and 72.75%, respectively. The Span-based model achieves the lowest F1 score (36.84%). It depends on dependency parsing results from the Stanford CoreNLP toolkit (Manning et al., 2014). Enhancing the performance of the Span-based model further necessitates an examination of appropriate dependency parsers tailored for the materials science domain.

In the PolymerAbstracts corpus (Shetty et al., 2023), our best-performing NER system, employing the BARTNER architecture, achieved an F1 score of only 67.57% for eight entity types. This indicates the enhanced consistency and quality of our MatSciNERE corpus.

Table 2. Results for NER on test set.

Method	P	R	F1
Span-based (Li et al., 2021)	59.56	26.66	36.84
Transition-based (Dai et al., 2020)	73.49	72.03	72.75
MaxClique (Wang et al., 2021)	77.35	71.86	74.51
BARTNER (Yan et al., 2021)	74.80	73.25	74.02
W2NER (Li et al., 2022)	**77.78**	**73.55**	**75.61**

Using the best trained model following the W2NER method (Li et al., 2022), we investigate the NER performance across entity types. The results are shown in Table 3.

The F1 scores demonstrate robust performance in identifying entities such as POLYMER, PROP_NAME, PROP_VALUE, MATERIAL_AMOUNT, and CHAR_METHOD. However, scores are lower for entities like CONDITION, COMPOSITE, OTHER_MATERIAL and REF_EXP, likely due to the varied contextual expressions and the annotating challenges encountered across the corpus. The F1 scores for ORGANIC and INORGANIC show significant enhancements in our NER module, with scores of 55.83% and 74.67%, respectively. By comparison, Shetty

et al. (2023) reported scores of only 26.2% for ORGANIC and 49.6% for INORGANIC.

Table 3. NER performance across types of entities.

Entity Type	F1	Entity Type	F1
POLYMER	83.12	PROP_NAME	80.00
MONOMER	74.52	PROP_VALUE	84.67
POLYMER_FAMILY	69.41	MATERIAL_AMOUNT	86.00
ORGANIC	55.83	CONDITION	63.97
INORGANIC	74.67	SYN_METHOD	74.16
COMPOSITE	61.70	CHAR_METHOD	90.80
OTHER_MATERIAL	32.65	REF_EXP	60.50

We also experimented with alternative encoders to enhance NER performance. Table 4 demonstrates that MatSciBERT (Gupta et al., 2022) yields better performance compared to the use of SciBERT (Beltagy et al., 2019).

Table 4. Results for NER on test set (other encoders).

Method	Pre-trained Model	P	R	F1
MaxClique (Wang et al., 2021)	BERT-large	77.35	71.86	74.51
	SciBERT	**80.64**	71.96	76.05
	MatSciBERT	78.65	75.50	77.04
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**

3.2 RE Performance

We compare the following models: (1) ATLOP (Zhou et al., 2021), a document-level RE (DocRE) model which aggregates contextual information by the Transformer attentions and adopts an adaptive threshold for different entity pairs, (2) DocuNet (Zhang et al., 2021), which models DocRE as a semantic segmentation task. We use the implementations of these models and apply to our data, and mostly follow the hyperparameters used in these models.

We define the RE task on our MatSciNERE corpus as a DocRE problem, where the gold entities are given in advance. An entity can have multiple mentions within the abstract, and a relation between two entities (e1, e2) exists if it is expressed by any pair of their mentions. During the inference step, the target is to predict relations between all possible entity pairs.

We also employ a rule-based approach to extract relations between two entity mentions and experimentally determined that the optimal distance between these mentions is within 70 characters. The rule-based method is applied to six relations: *has_property, has_value, has_amount, has_condition, synthesized_by,* and *characterized_by*. The type constraints for each head and tail entity mention are derived from our proposed relation schema, as depicted in Figure 3.

In the case of ATLOP and DocuNet models, we use different encoders: BERT-large, SciBERT, MatSciBERT and RoBERTa (Liu et al., 2019, one of original encoders for the DocuNet model) and then evaluate the performance results for RE.

As illustrated in Table 5, ATLOP achieves a highest F1 score of 83.23%, while DocuNet achieves 82.34%. Notably, the ATLOP model with MatSciBERT achieves the highest performance. Conversely, DocuNet shows lower scores when other encoders are utilized, even worse compared to the same model using BERT-large.

Table 5. Results for RE on test set given gold entities.

Method	Pre-trained Model	P	R	F1
Rule-based RE	-	33.67	39.94	36.54
ATLOP (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**
DocuNet (Zhang et al., 2021)	BERT-large	79.18	**85.76**	82.34
	SciBERT	74.93	78.04	76.45
	MatSciBERT	75.57	78.04	76.79
	RoBERTa	69.91	70.33	70.12

Moreover, the rule-based method only achieves an F1 score of 36.54%, while ATLOP and DocuNet significantly improved upon this. This represents a substantial enhancement in performance when transitioning from the rule-based method to the automated DocRE methods.

To assess the practical applicability and robustness of our trained NER and RE models, we choose an additional set of 250 paragraphs from diverse materials science papers, chosen from a large selection of thousands of paragraph candidates. These polymer paragraphs serve as the input for evaluating the performance of our top-performing NER and RE systems.

More specifically, we employ the trained $W2NER_{MatSciBERT}$ model to identify entity mentions in each sentence of the input paragraph. These predicted entity mentions are then aggregated into the corresponding abstract. Subsequently, we utilize the trained $ATLOP_{MatSciBERT}$ model to extract relations between pairs of entities at the document level.

We also apply the same annotation process (outlined in Section 2.3) to the above 250 materials science paragraphs. The ratio of each entity type as well as relation type is shown in Figure 5. We present our RE system performance in an end-to-end setting in Table 6.

Our NER module demonstrates robust performance with an F1 score of 84.97%. However, there remains a gap in RE performance between gold entities settings (83.23%) and end-to-end settings (65.54%), indicating the need for further improvement in our RE module. Nevertheless, we believe that prior research has not provided such a high number of relation annotations for both general material science and its subdomain like polymer science. Additionally, practical NER and RE systems in these domains are very limited.

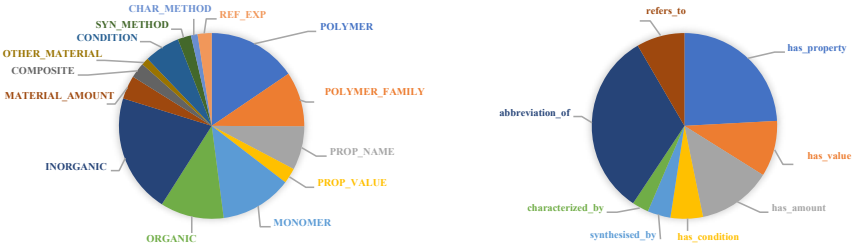


Fig. 5. Annotations on 250 paragraphs.

Table 6. Performance of our end-to-end RE system on 250 paragraphs.

Module	P	R	F1
NER	86.09	83.88	84.97
RE	65.15	65.95	65.54

4 Related Work

4.1 Resources for NER and RE in Materials Science Domain

The availability of resources for hierarchical ontology concerning entities in the broader field of materials science, including specific subdomains, is very limited for both entity and relation annotations. Despite the prevalence of fundamental entity categories such as organic, inorganic, and polymers, comprehensive annotations for all these entity types in materials science text are currently lacking to the best of our knowledge.

Isazawa et al. (2022) referenced two NER datasets, namely CHEMDNER (Krallinger et al., 2015) and Matscholar (Weston et al., 2019), suggesting that a system could be trained on a combination of sentences from both organic (primarily from CHEMDNER) and inorganic (mostly from Matscholar) chemistry. However, our investigation revealed the absence of separate annotations for organic or inorganic entities in these datasets. Consequently, a NER model trained on such data may be significantly influenced by incorrect labels.

Mysore et al.'s (2019) dataset comprises 230 labeled synthesis procedures specifically for inorganic synthesis. Yamaguchi et al. (2020) superconductive materials for the NER task. O'Gorman et al. (2021) released the largest NER dataset for procedural text in materials science. Yang et al. (2022) introduced PcMSP, a corpus for

both NER and RE from polycrystalline materials synthesis procedures. Shetty et al. (2023) released 750 abstracts of polymers for the NER task, containing eight distinct entity types.

4.2 Methods in NER and RE

Existing NER methods are primarily categorized into sequence labeling-based (Huang et al., 2015; Lample et al., 2016; Chiu and Nichols, 2016), span-based (Luan et al., 2019; Shen et al., 2021), and generation-based (Strakova et al., 2019; Paolini et al., 2021; Yan et al., 2021) methods. Sequence labeling approaches have limitations in handling entities with multiple labels that overlap. Span-level classification is more adept at handling overlapped entities but may face challenges with a high number of entities. Generative language model-based methods provide an alternative perspective, treating NER tasks as problems of generating sequences of entity spans. This approach can effectively handle flat, overlapped, or discontinuous entity mentions. Additionally, there are other works that specifically focus on discontinuous NER.

In relation extraction models, a common approach involves employing a pipeline with RE following NER (Huang et al., 2021). An alternative strategy is the joint entity and RE (Giorgi et al., 2022; Lu et al., 2022), where models are designed to simultaneously perform both tasks on a given text.

5 Conclusion

In this work, we introduce MatSciNERE, a comprehensive corpus for materials science comprising named entities and relations, developed from a previous version focused on polymer science. Our pipeline system, emphasizing practicality, features NER and RE modules adept at handling complex expressions and diverse contexts. We plan to release the MatSciNERE corpus, including 1,000 annotated abstracts and paragraphs, along with trained models, to support research in material science and polymer science.

References

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. Conference on Empirical Methods in Natural Language Processing.

Straková, J., Straka, M., & Hajic, J. (2019). Neural Architectures for Nested NER through Linearization. ArXiv, abs/1908.06926.

Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., Sun, L., & Wu, H. (2022). Unified Structure Generation for Universal Information Extraction. Annual Meeting of the Association for Computational Linguistics.

Giorgi, J., Bader, G.D., & Wang, B. (2022). A sequence-to-sequence approach for document-level relation extraction. Workshop on Biomedical Natural Language Processing.

Huang, K., Tang, S., & Peng, N. (2021). Document-level Entity-based Extraction as Template Generation. ArXiv, abs/2109.04901.

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). A Unified Generative Framework for Various NER Subtasks. ArXiv, abs/2106.01223.

Paolini, G., Athiwaratkun, B., Krone, J., Ma, J., Achille, A., Anubhai, R., Santos, C.N., Xiang, B., & Soatto, S. (2021). Structured Prediction as Translation between Augmented Natural Languages. ArXiv, abs/2101.05779.

Luan, Y., Wadden, D., He, L., Shah, A., Ostendorf, M., & Hajishirzi, H. (2019). A general framework for information extraction using dynamic span graphs. North American Chapter of the Association for Computational Linguistics.

Shen, Y., Ma, X., Tan, Z., Zhang, S., Wang, W., & Lu, W. (2021). Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition. Annual Meeting of the Association for Computational Linguistics.
Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. ArXiv, abs/1508.01991.
Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. North American Chapter of the Association for Computational Linguistics.
Chiu, J.P., & Nichols, E. (2015). Named Entity Recognition with Bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 4, 357-370.
Mysore, S., Jensen, Z., Kim, E.J., Huang, K., Chang, H., Strubell, E., Flanigan, J., McCallum, A., & Olivetti, E.A. (2019). The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. LAW@ACL.
Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O.V., Trewartha, A., Persson, K.A., Ceder, G., & Jain, A. (2019). Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. Journal of chemical information and modeling.
Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki. 2020. SC-CoMIcs: A Superconductivity Corpus for Materials Informatics. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6753–6760, Marseille, France. European Language Resources Association.
O'Gorman, T.J., Jensen, Z., Mysore, S., Huang, K., Mahbub, R., Olivetti, E.A., & McCallum, A. (2021). MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text. Conference on Empirical Methods in Natural Language Processing.
Yang, X., Zhuo, Y., Zuo, J., Zhang, X., Wilson, S., & Petzold, L. (2022). PcMSP: A Dataset for Scientific Action Graphs Extraction from Polycrystalline Materials Synthesis Procedure Text. ArXiv, abs/2210.12401.
Shetty, P., Rajan, A.C., Kuenneth, C., Gupta, S., Panchumarti, L.P., Holm, L., Zhang, C., & Ramprasad, R. (2023). A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. Npj Computational Materials, 9.
Isazawa, Taketomo, and Jacqueline M. Cole. "Single model for organic and inorganic chemical named entity recognition in ChemDataExtractor." Journal of Chemical Information and Modeling 62.5 (2022): 1207-1213.
Krallinger, Martin, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman et al. "The CHEMDNER corpus of chemicals and drugs and its annotation principles." Journal of cheminformatics 7, no. 1 (2015): 1-17.
Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. 2021. A Span-Based Model for Joint Overlapped and Discontinuous Named Entity Recognition. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4814–4828, Online. Association for Computational Linguistics.
Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An Effective Transition-based Model for Discontinuous NER. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5860–5870, Online. Association for Computational Linguistics.
Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021. Discontinuous Named Entity Recognition as Maximal Clique Discovery. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 764–774, Online. Association for Computational Linguistics.
Li, Jingye, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. "Unified named entity recognition as word-word relation classification." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 10, pp. 10965-10973. 2022.
Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.
Kingma, D.P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. CoRR, abs/1412.6980.
Gupta, Tanishq, Mohd Zaki, NM Anoop Krishnan, and Mausam. "MatSciBERT: A materials domain language model for text mining and information extraction." npj Computational Materials 8, no. 1 (2022): 102.
Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).