PolyNERE: A Novel Ontology and Corpus for Named Entity Recognition and Relation Extraction in Polymer Science Domain

Van-Thuy Phi[1], **Hiroki Teranishi**[1], **Yuji Matsumoto**[1], **Hiroyuki Oka**[2], **Masashi Ishii**[2]
[1]RIKEN Center for Advanced Intelligence Project
[2]Center for Basic Research on Materials, National Institute for Materials Science
{thuy.phi, hiroki.teranishi, yuji.matsumoto}@riken.jp, {oka.hiroyuki, ishii.masashi}@nims.go.jp

Abstract

Polymers are widely used in diverse fields, and the demand for efficient methods to extract and organize information about them is increasing. An automated approach that utilizes machine learning can accurately extract relevant information from scientific papers, providing a promising solution for automating information extraction using annotated training data. In this paper, we introduce a polymer-relevant ontology featuring crucial entities and relations to enhance information extraction in the polymer science field. Our ontology is customizable to adapt to specific research needs. We present PolyNERE, a high-quality named entity recognition (NER) and relation extraction (RE) corpus comprising 750 polymer abstracts annotated using our ontology. Distinctive features of PolyNERE include multiple entity types, relation categories, support for various NER settings, and the ability to assert entities and relations at different levels. PolyNERE also facilitates reasoning in the RE task through supporting evidence. While our experiments with recent advanced methods achieved promising results, challenges persist in adapting NER and RE from abstracts to full-text paragraphs. This emphasizes the need for robust information extraction systems in the polymer domain, making our corpus a valuable benchmark for future developments.

Keywords: Polymer Corpus, Named Entity Recognition, Relation Extraction, Property Information Extraction

1. Introduction

With the increasing use of polymers in various fields, there is a growing need for efficient methods to collect and organize information about them. One important source of information about polymers is scientific papers in the materials domain. These papers contain information about newly introduced polymers' properties, synthesis methods, etc. However, tracking relevant information about polymer entities can be challenging due to the vast amount of data.

Recent advancements in natural language processing (NLP) and machine learning have enabled the development of automated methods for extracting information from scientific papers. A named entity recognition (NER) and relation extraction (RE) system can be used to automatically recognize important terms (entities) or groups of terms/expressions that align with specific categories in the data (NER), and it can extract relationships between these entities (RE). In the context of polymer research, automated methods can extract new polymer-relevant names, property names and values of existing materials and other useful information from scientific papers, thus reducing the effort of manual extraction. As a practical application, the extracted data can be used to enhance polymer databases such as PoLyInfo[1] (Otsuka et al, 2011), which currently depends on manual updates by human experts.

Nonetheless, the development of NER and RE systems is applicable if manually annotated corpora are available. These corpora allow for training and evaluating NER and RE models. In addition, one of the main challenges in using machine learning for information extraction in polymer science is the complex and diverse language used to describe polymer entities and their properties. For example, polymers can be referred to by different names or abbreviations, and properties can be described using a range of terminology[2]. This complexity makes it difficult for automated methods to accurately extract information. To address these challenges and accelerate the research in polymer science, there is an urgent need for a substantial, high-quality manually annotated dataset that encompasses rich information, covering not only polymers but also essential related entities and their relationships.

In our study, we aim to develop a corpus that confronts the gap in information extraction from polymer science literature. Specifically, our focus is on simultaneous extraction of entities (polymers and relevant materials), relations (involving associations between materials and their corresponding property names and values, etc.) and other supporting relations (including abbreviations and coreference).

Our contributions are as follows. First, we present an ontology comprising key polymer-related entities and relationships, emphasizing fundamental concepts in the polymer domain, while also enabling a focus on specific entities and relations of interest[3]. This ontology's flexibility enables customization to improve information extraction systems within the field of polymers.

We have developed PolyNERE, a newly created and high-quality corpus for NER and RE. It consists of 750 polymer paper abstracts with text sources closely matching the PolymerAbstracts corpus (Shetty et al. 2023). Each includes raw text and DOI information, annotated as per our ontology.

Our corpus possesses several distinctive features: (i) fourteen types of entities and eight types of relations,

[1] https://polymer.nims.go.jp/
[2] See Appendix A for the challenges of the polymer science domain.
[3] This ontology differs from the PoLyInfo ontology and conceptual schema (https://doi.org/10.48505/nims.4413), which aims to systematize polymer chemistry.

LREC-COLING 2024, pages 12856–12866

Entity Type	Definition	Example
POLYMER	Material entities that are polymers	*"Sulfonated poly(phthalazinone ether ketone nitrile)", "polyethylene"*
POLYMER_FAMILY	Material entities that refer to a class of polymers	*"bio-polyimides", "PIs", "epoxy", "poly(amic acid)s", "polyanhydride"*
PROP_NAME	Entity type which indicates a specific material property	*"ion conductivity", "power density", "glass transition temperature"*
PROP_VALUE	Entity type which includes a numeric value and its unit for a specific material property	*"9400 g/mol", "less than 16 wt%", "> 100,000 g/mol"*
MONOMER	Material entities which are explicitly indicated as being the repeat units for a POLYMER entity	*"N-isopropylacrylamide", "4,4'-bisphenol"*
ORGANIC	Material entities that are organic but not polymers	*"hydroxy urea", "divinyl benzene", "maleic acid", "PFSA"*
INORGANIC	Material entities which are inorganic and are typically used as additives in a polymer formulation	*"Ag", "indium(III) oxide", "In2O3"*
MATERIAL_AMOUNT	Entity type which indicates the amount of a particular material in a material formulation	*"90%", "5 wt.%", "10 mass%"*
COMPOSITE	A material formed by combining two or more distinct components to achieve improved properties not present in the individual constituents	*"TiO2-DA-PEI", "GO/PVA", "PVdF:PEMA"*
OTHER_MATERIAL	Materials entities that do not fall under the specific entity types mentioned above. They can be described without chemical specificity, or can indicate other materials besides existing entity types, including mixtures, etc.	*"anion exchange membranes", "ethanol/water", "porous film"*
CONDITION	A condition in which the value of the property is measured	*"at 50°C", "using air O2", "between 15 and 60°C", "with increasing Mn"*
SYN_METHOD	Any technique for synthesising a material	*"ring-opening polymerization", "radical terpolymerization"*
CHAR_METHOD	Any method used to characterize a material	*"dynamic light scattering", "Neutron transmission measurements"*
REF_EXP	Short for referring expression, i.e., a phrase which is usually used to refer to an entity in the previous context	*"They", "this polymer", "the resulting copolymers", "the materials", "Its"*

Table 1: Definitions of annotated entity types according to our ontology

allowing for the effective handling of intricate contextual scenarios, particularly with the inclusion of special relations. (ii) Various settings for NER, encompassing flat, overlapped, and discontinuous mentions. (iii) Annotated entities and relations can be asserted at the sentence- or document-level. (iv) It provides reasoning capabilities for the RE task through supporting evidence derived from our two supporting relations. We expect that the corpus will serve as a benchmark for advancing information extraction in the polymer domain. We plan to release our corpus to the research community.

To assess the challenges of PolyNERE, we adopt a variety of recent advanced NER and RE methods under various settings. The experiments show that, despite promising results in extracting our target entities and relations, there is considerable room for improvement in general-purpose NER and RE systems. Additionally, our error analysis in extracting N-ary relation tuples, especially in unseen material paragraphs, highlights the challenges faced in adapting NER and RE systems from abstracts to full-text paragraphs, emphasizing the need for robust information extraction systems in the polymer domain.

2. Polymer Corpus Construction

Our goal is to construct a polymer corpus for abstract-level named entity recognition (NER) and relation extraction (RE) from plain text, featuring high-quality annotations and rich information content. The dataset captures essential information about polymer-related entities and their relations frequently found in polymer and materials abstracts.

2.1 Data Source

To construct our corpus, we start with the collected abstracts in the PolymerAbstracts corpus (Shetty et al., 2023). The texts were obtained from APIs and websites of publishers such as Elsevier, Wiley, etc. 750 abstracts from this corpus were annotated and utilized to train NER models. Unlike some other studies, Shetty et al. (2023) did not employ the BIO tagging scheme, which denotes the Beginning-Inside-Outside of the labeled entity. Instead, a simpler IO labeling approach was employed where only tokens belonging to target entities are annotated while all other tokens are labeled as "OTHER". Consequently, only flat entity mentions can be annotated. In addition,

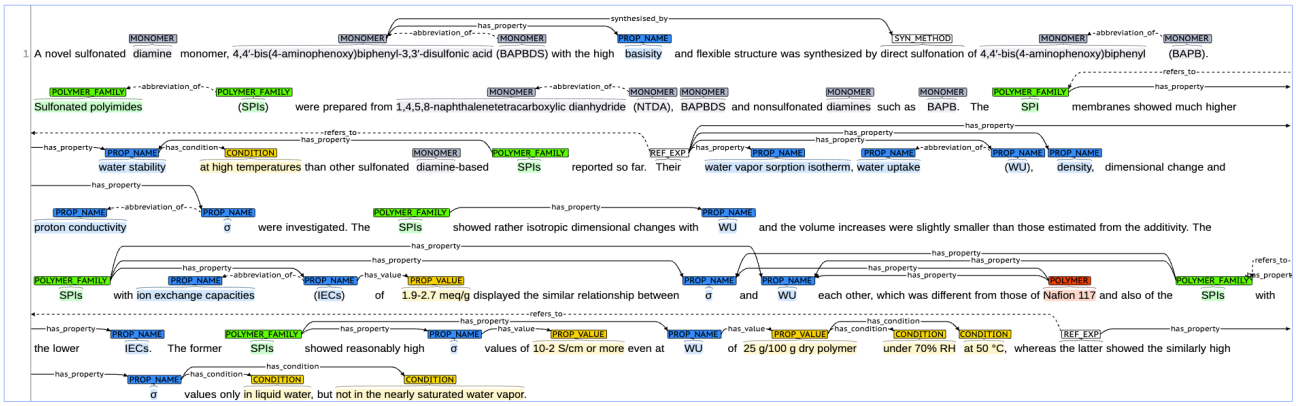

Figure 1: Annotation sample

annotations were supplied for abstract texts that were tokenized. This may impact the performance of automated NER systems.

Initially, these collected data were meant to recognize named entities. However, in this work, we use them for the purpose of developing a NER and RE corpus. The annotations in our dataset are derived from manually collected raw texts, drawing from similar sources found in PolymerAbstracts. Each raw abstract contains DOI information, potentially aiding in alignment with other components of papers, such as full texts, titles, tables, etc. Moreover, our raw text version closely aligns with the original abstract texts. For instance, in terms of property values, we use "*0.2 cm3 m-2·day-1*" as the text, instead of "*0.2 cm^{3 } m^{- 2}*day^{-1 }*" in PolymerAbstracts.

2.2 Entity Annotations

Our dataset provides annotations for various types of polymer-relevant materials, which were categorized primarily based on their usefulness for downstream polymer science tasks. We define a total of fourteen (14) entity types, which are detailed in Table 1.

We retained similar entities such as POLYMER, POLYMER_FAMILY, MONOMER, ORGANIC, INORGANIC, and MATERIAL_AMOUNT as defined in (Shetty et al., 2023). However, we made slight adjustments to the definitions of PROP_NAME and PROP_VALUE to focus on specific property details. Additionally, we developed an entity ontology containing concepts relevant to polymers. Our ontology incorporates other novel entity types, namely COMPOSITE, OTHER_MATERIAL, CONDITION, SYN_METHOD, CHAR_METHOD, and REF_EXP.

Figure 1 shows an example of entity annotations highlighted and visualized with BRAT (Stenetorp et al., 2012) annotation tool.

For a clearer understanding of the ontology of all entities, please refer to 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.

While our primary focus is on entities and relations related to polymers and their associated information, we also included entities like organic and inorganic, which are commonly found in composites or other material types. These additional entities play a crucial role in distinguishing polymers from other potentially confusing entities in the polymer domain. Moreover, our ontology can serve as a foundation for extending the development of general-purpose information extraction systems.

For entity annotations, we focus on: (i) Annotate the most specific material mentions, e.g., the complete phrase "*sulfonated poly(ether ether ketone)*" should be preferred over "*poly(ether ether ketone)*" as it is more specific. (ii) Annotate minimum necessary text spans for an entity, i.e., without brackets, punctuation marks, etc. For example, for a PROP_VALUE (property value), we choose "*120 °C*" instead of "*120 °C.*" which contains a redundant '.' character, or for a PROP_NAME (property name), we choose "*Tg*" instead of "*(Tg)*" with a redundant pair of brackets.

2.3 Relation Annotations

In our ontology, MATERIAL_GROUP is defined as the group that contains the following material entities: POLYMER, POLYMER_FAMILY, MONOMER, ORGANIC, INORGANIC, COMPOSITE, and OTHER_MATERIAL.

We define eight (8) relation types as illustrated in Figure 2:

has_property: Indicates that the MATERIAL_GROUP entity possesses or exhibits the property described by the PROP_NAME entity.

has_value: Describes the relation between a specific property name (PROP_NAME) and its corresponding property value (PROP_VALUE).

When PROP_NAME is not present or lacks clarity in describing the association between the <PROP_NAME, PROP_VALUE> pair, it becomes essential to annotate the same relationship label '*has_value*' to the <MATERIAL_GROUP, PROP_VALUE> pair.

has_amount: Indicates a relation between a MATERIAL_GROUP entity and a MATERIAL_AMOUNT entity, indicating the proportion or quantity of the material involved.

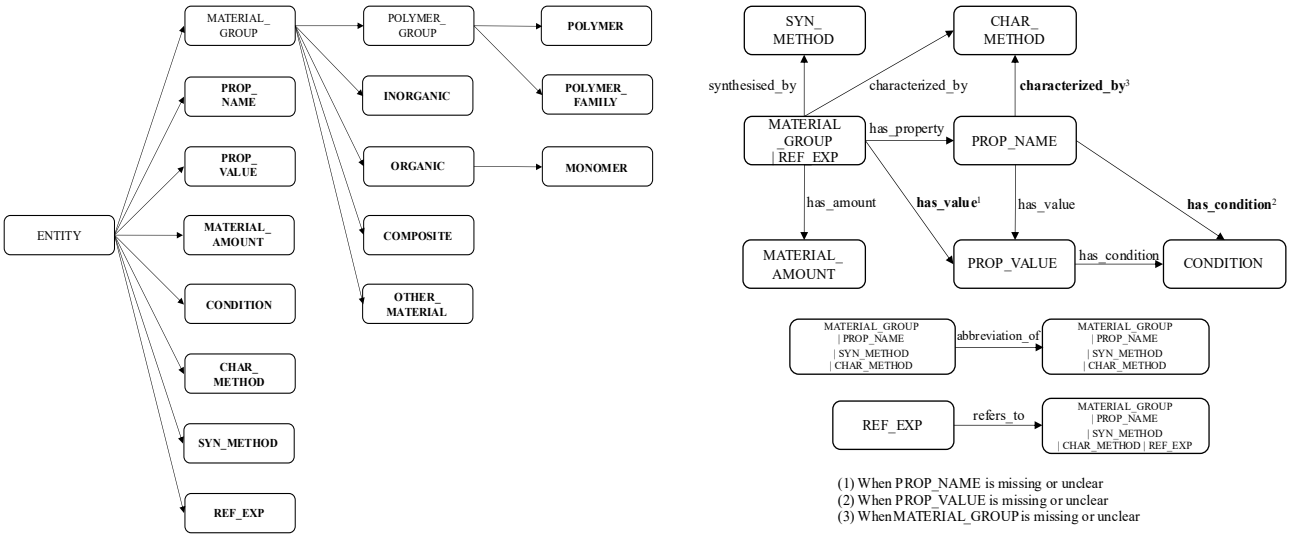

Figure 2: Left: Ontology of entities separates types by their definitions; monomers are organic; polymers can be inorganic or organic; Right: Illustration of material entity relationships

has_condition: Captures a relation between a property value (PROP_VALUE) and a condition (CONDITION). A single property value may have multiple conditions.
When PROP_VALUE is not present or lacks clarity in describing the association between the <PROP_VALUE, CONDITION> pair, it becomes essential to annotate the same relationship label '*has_condition*' to the <PROP_NAME, CONDITION> pair.

synthesised_by: Indicates that the SYN_METHOD entity is the particular synthesis method used in the production of the MATERIAL_GROUP entity.

characterized_by: Indicates that the CHAR_METHOD entity is a particular analytical method or test utilized to understand and describe the properties and characteristics of the MATERIAL_GROUP entity.
When MATERIAL_GROUP is not present or lacks clarity in describing the association between the <MATERIAL_GROUP, CHAR_METHOD> pair, it becomes essential to annotate the same relationship label '*characterized_by*' to the <PROP_NAME, CHAR_METHOD> pair.

abbreviation_of: Indicates a relation between two terms, where one term is an abbreviation or acronym of the other term. Abbreviation pairs are applicable not only to MATERIAL_GROUP entities, but also to PROP_NAME, SYN_METHOD and CHAR_METHOD.

refers_to: Indicates a relation between two terms that refer to the same entity or concept in a text. Our aim is to minimize the distance between two arguments involved in this relation. Consequently, the *'refers_to*' relation is most desirable when connecting the closest entity mentions. The key objective is to establish connections between the most prominent REF_EXP mentions, both within and across sentences, to offer detailed insights into coreference.

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.
Figure 1 illustrates relation annotations that have been visualized using the BRAT tool, relying on the previously introduced entity ontology.
For relation annotations, we focus on: (i) Properly annotate the relation pairs with their corresponding directions. (ii) annotated relations which are explicitly mentioned in the text. (iii) three (3) special cases when the context is either unclear or missing. For accurate depictions of relationship directions and special relations (i.e., '*has_value*', '*has_condition*' and '*characterized_by*'), please refer to Figure 2.

2.4 Annotation Process

In this section, we elaborate on our thorough annotation process, following the new annotation assumption and guided by our entity and relation ontology, thus establishing PolyNERE as a novel and unique corpus when compared to existing corpora.
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 PolyNERE 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:

#abstracts	750
#sentences/abstract	15.25
#tokens/sentence	11.87
#entities/abstract	25.24
#relations/abstract	15.29
Overlapped entities	2,148 mentions
Discontinuous entities	269 mentions
ENTITY (14)	Total: 18,930 mentions
POLYMER	3,988 (577/750 abstracts)
POLYMER_FAMILY	1,145 (308)
PROP_NAME	3,823 (715)
PROP_VALUE	1,815 (586)
MONOMER	1,470 (311)
ORGANIC	617 (158)
INORGANIC	908 (202)
MATERIAL_AMOUNT	485 (238)
COMPOSITE	392 (171)
OTHER_MATERIAL	175 (85)
CONDITION	717 (351)
SYN_METHOD	378 (231)
CHAR_METHOD	1,747 (433)
REF_EXP	1,270 (459)
RELATION (8 +3 special)	Total: 11,471 pairs
has_property	3,447 (660/750 abstracts)
has_value	1,879 (581)
has_amount	349 (185)
has_condition	970 (345)
synthesised_by	278 (193)
characterized_by	1,343 (389)
abbreviation_of	1,869 (617)
refers_to	1,336 (458)

Table 2: Corpus statistics

1st round: In the first 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. Under our new annotation assumption, we have expanded the scope of material entities. Previously, only material entities explicitly linked to a <PROP_NAME, PROP_VALUE> pair in the abstract were labeled (Shetty et al., 2023). However, recognizing the potential impact on NER and RE models, we aim to include as many entity mentions and their relations as possible within our coverage. Subsequently, we manually check and add missing annotations in 750 abstracts. Our relation schema captures detailed information about the relationships between polymer-related entities.

Following the annotation of the eight (8) aforementioned entity types, we observe that 44.07% of mentions in the PolymerAbstracts corpus have been altered in our PolyNERE corpus, signifying a substantial degree of modification.

Furthermore, we carry out annotations for the six (6) new entity types incorporated into our ontology.

2nd round: In the second round, our emphasis is on annotating eight (8) types of relations, along with three (3) 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: In the third 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, the refined annotation guidelines and consistent annotations are enhanced through ongoing discussions between the annotator and a polymer expert. The annotator seeks guidance from the polymer expert when necessary, with revisions are primarily retained as the final version after the third round.

3. Corpus Statistics

Our PolyNERE corpus consists of a total of 750 abstracts, divided into three sets: 637 for training, 38 for development, and 75 for testing. The maximum number of entities and relations per abstract is 76 and 79, respectively.

Table 2 displays the statistics for our corpus, presenting details about the annotation type, and the number of annotations across various categories within PolyNERE. Overall, our PolyNERE 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.

PolyNERE contains 18,930 entity mentions, which is 1.74 times higher than the number found in PolymerAbstracts corpus. There are 2,148 overlapped entity mentions, constituting 11.35% of all entity mentions. The number of discontinuous entity mentions is 269, representing 1.42% 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 11,471, and to the best of our knowledge, none of the prior works in polymer research have incorporated such a large number of relation annotations.

To assess the quality of the corpus, we randomly selected 10 polymer 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.

Type	Method	P	R	F1
Sequence Labeling	**LinearedCRF** (Strakova et al., 2019)	75.93	69.79	72.73
	Second-best (Shibuya and Hovy, 2020)	75.78	72.02	73.85
Span-based	**Biaffine** (Yu et al., 2020)	**78.37**	70.86	74.42
	Pyramid (Wang et al., 2020)	73.50	71.43	72.45
MRC-based	**MRC** (Li et al., 2020)	77.52	68.78	72.89
Generation-based	**Seq2seq** (Strakova et al., 2019)	75.56	71.30	73.37
	BARTNER (Yan et al., 2021)	74.20	**75.86**	**75.02**

Table 3: Performance of NER models on test set

4. Experiments

4.1 Experimental Settings

To assess the challenges of the PolyNERE corpus, we conducted experiments to show the performance of recent advanced NER and RE models.

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 a given abstract. Given the low ratio of discontinuous mentions (1.42%), we exclude them from the evaluation. Specifically, we only consider flat and overlapped mentions, while discontinuous ones are omitted.

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

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 consisting of 75 abstracts.

4.2 NER Performance

We use the following models as baselines: (1) Sequence labeling-based models: LinearedCRF (Strakova et al., 2019), Second-best (Shibuya and Hovy, 2020), (2) Span-based models: Biaffine (Yu et al., 2020), Pyramid (Wang et al., 2020), (3) MRC-based model (Li et al., 2020), (4) Generation-based models: Seq2seq (Strakova et al., 2019), BARTNER (Yan et al., 2021).

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 3 shows the evaluation of NER on the test set. The BARTNER model achieves the best performance on PolyNERE, achieving the highest recall (75.86) and F1 score (75.02) in comparison to other approaches. It outperforms the previous best F1 score, represented by the Biaffine model, by a margin of 0.6%. The Biaffine model also achieved a highest precision of 78.37%. The MRC model obtains a lower F1 score on the PolyNERE dataset. In fact, the MRC model heavily relies on the definitions of entity types for constructing queries to extract semantic relations between entities. Further improvement in MRC model's performance requires a thorough investigation of entity definitions.

The experimental results demonstrate the advantages of generation-based and span-based methods over sequence labeling-based approaches on our PolyNERE corpus, with the highest F1 score reaching 75.02%. This aligns with recent developments in NER for material datasets, such as SC-CoMIcs (Yamaguchi et al., 2020). In the PolymerAbstracts corpus (Shetty et al., 2023), our best-performing NER system, utilizing the same BARTNER architecture, obtains an F1 score of 67.57 for eight (8) entity types. This indicates the improved consistency and quality of our PolyNERE corpus.

Entity Type	P	R	F1
POLYMER	84.17	79.76	81.91
POLYMER_FAMILY	59.68	69.81	64.35
MATERIAL_GROUP	73.58	69.86	71.67
PROP_NAME	83.25	82.19	82.72
PROP_VALUE	78.91	84.67	81.69
MATERIAL_AMOUNT	72.34	77.27	74.72
CONDITION	53.52	46.91	50.00
SYN_METHOD	64.58	86.11	73.81
CHAR_METHOD	87.72	93.75	90.63
REF_EXP	60.66	60.16	60.41

Table 4: NER performance across entity types

Using the best trained model following the BARTNER method, we investigate the NER performance across entity types. The results are shown in Table 4.

Due to our specific focus on the polymer science domain, we only showed scores for ten entity types primarily related to polymers. The F1 scores for POLYMER, PROP_NAME, and PROP_VALUE are 81.91%, 82.72%, and 81.69%, respectively, demonstrating relatively strong performance for our primary entities of interest at this stage. However, the F1 score for recognizing POLYMER_FAMILY is 64.35%, revealing challenges in classifying entity mentions as either POLYMER or POLYMER_FAMILY.

Also, the BARTNER model performs worse for other entities like CONDITION and REF_EXP, potentially

due to the diverse contextual expressions of these entity mentions and the difficulty in annotating them throughout the corpus. In the case of MATERIAL_GROUP, the F1 score of 71.67% suggests substantial potential for enhancement in NER systems.

Method	Pre-trained Model	P	R	F1
Rule-based	-	31.77	52.46	39.57
ATLOP (Zhou et al., 2021)	BERT-base	77.60	79.10	78.34
	BERT-large	82.20	81.67	81.93
	SciBERT	82.44	**89.07**	**85.63**
DocuNet (Zhang et al., 2021)	BERT-base	80.74	76.85	78.75
	BERT-large	**84.44**	85.53	84.98
	SciBERT	83.96	79.10	81.46

Table 5: Performance of RE models on test set

4.3 RE Performance

We use the following models as baselines: (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 baseline models and apply to our polymer data, and mostly follow the hyperparameters used in the baseline models.

We define the RE task on our PolyNERE 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 two sentences. 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 2.

In the case of ATLOP and DocuNet models, we use three different encoders: BERT-base, BERT-large, and SciBERT (Beltagy et al., 2019), and then report the performance results for RE. We run each model 5 times and use the deveulopment data to pick the best model.

As shown in Table 5, when employing the BERT-base encoder, the results are quite comparable, with F1 scores of 78.34 and 78.75 for ATLOP and DocuNet, respectively. However, for models based on BERT-large, significantly better results are achieved, yielding an F1 score of 81.93% for ATLOP and 84.98% for DocuNet, which is the highest score observed for the latter. On the contrary, using the SciBERT encoder reveals the opposite trend: ATLOP reaches the highest F1 score (85.63), while DocuNet obtains a lower score of 81.46, even falling below the performance of the same model based on BERT-large.

Moreover, the rule-based method achieved an F1 score of 39.57%, while ATLOP and DocuNet significantly improved upon this, achieving an F1 score of 85.63% and 84.98%, respectively. This represents a substantial enhancement in performance when transitioning from the rule-based approach to the automated DocRE methods such as ATLOP and DocuNet.

4.4 Applications Involving Polymer Property Information Extraction

To assess the practical applicability and robustness of our trained NER and RE models, we choose an additional set of 250 paragraphs from diverse material papers, primarily featuring the '*poly*' prefix in their abstracts. They have been included in the papers under our licensing agreement with publishers, including Elsevier, the American Chemical Society, and others. We converted the XML files to plain text files using our processing script. These polymer paragraphs[4] serve as the input for evaluating the performance of our top-performing NER and RE systems, as outlined in sections 4.2 and 4.3.

More specifically, we employ the trained $\text{BARTNER}_{\text{BART-large}}$ 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 $\text{ATLOP}_{\text{BERT-large}}$ model to extract relations between pairs of entities at the document level. We choose $\text{ATLOP}_{\text{BERT-large}}$ for the RE model to align with the BERT-large encoder utilized in most of our NER models.

At the current stage of our research, our primary focus is on developing an end-to-end practical RE system, with a specific emphasis on the polymer entity and its associated property information. Therefore, based on the output of the ATLOP model, we retain only the following relations, *'has_property*', '*has_value*', and '*refers_to*', which involve four entity types, POLYMER, PROP_NAME, PROP_VALUE and REF_EXP.

By utilizing the predicted entity mentions and relation pairs within each polymer paragraph, we can derive tuples related to polymer property information in the following format: t=<POLYMER | REF_EXP,

[4] The DOIs of papers containing these 250 paragraphs will be made openly accessible.

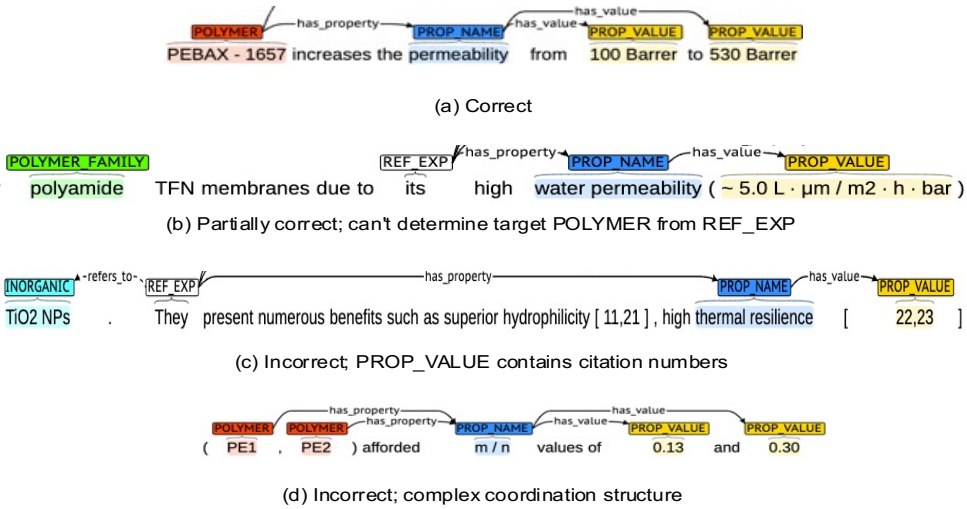

Figure 3: Examples of correct and incorrect predictions

PROP_NAME, PROP_VALUE1, PROP_VALUE2, …>. This is achieved by aggregating relation pairs that share common head or tail entities.

Each tuple t can be regarded as an n-ary relation mention, with n>=3. Specifically, each tuple t comprises exactly one POLYMER entity (or one REF_EXP entity), one PROP_NAME entity, and at least one PROP_VALUE entity.

A tuple, denoted as t = <POLYMER, PROP_NAME, PROP_VALUE1, PROP_VALUE2, …>, is deemed correctly extracted and receives a score of 1.0 if it accurately represents the true facts at the paragraph level. This entails the correct recognition of all entities within t, as well as the validation of the relations as true positives. If t takes the form t = <REF_EXP, PROP_NAME, PROP_VALUE1, PROP_VALUE2, …> and is correctly extracted, a score of 0.5 is assigned. In this scenario, the REF_EXP entity doesn't reference other entity mentions within the paragraph. Otherwise, a score of zero (0.0) is given. Then, we report the accuracy of all extracted tuples t. We conduct a manual evaluation to assess the accuracy of the extracted tuples, aiming to understand the difficulties encountered when transitioning NER and RE systems from abstracts to unseen paragraphs in full articles. In total, our NER and RE systems, trained on abstract texts, extracted 69 tuples related to polymer property information from 250 polymer paragraphs. Of these, 42 were deemed correct, resulting in an accuracy of 60.87%.

Figure 3 shows sample predictions produced by our NER and RE system: (a) The system extracts correctly for the POLYMER mention and its associated property name and values, (b) The system is not able to determine the target POLYMER from REF_EXP, i.e., there is no connection from '*Its*' to '*polyamide*', (c) The citation numbers are mispredicted as PROP_VALUE, therefore this tuple is regarded as incorrect, (d) Insufficient predictions for relation pairs, leading to not possible to infer the correct tuples. In addition, there are other situations that require reasoning from tables or figures or from other parts of the articles for accurate tuple extraction. All of the challenges outlined above point towards promising directions for our future work. This includes extending the dataset from abstracts to full paragraphs and enhancing the recognition of "REF_EXP" chains to better provide evidence and support reasoning about the target relations.

5. Related Work

5.1 Automatic Extraction of Entities and Relations

Extracting relevant named entities and relations in polymer research can present difficulties due to the complex terminology used (such as distinguishing between monomers and organic compounds, or between polymer families and individual polymers), the complex structures involved in describing various properties, and the various data formats used (such as the use of IUPAC nomenclature for naming scientific polymers). Accurate analysis and extraction in this domain require a deep understanding of polymer science, chemistry, and materials science.

The main categorization of existing NER methods includes 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 are limited in handling nested entities with multiple labels. Span-level classification is more suitable for handling nested entities, but it may struggle with a high number of entities. Generative language model-based approaches offer an alternative approach, treating NER tasks as entity span sequence generation problems.

Typically, relation extraction models are used in a pipeline after entity prediction (Huang et al., 2021). Alternatively, a joint entity and relation extraction approach can also be employed (Giorgi et al., 2022; Lu et al., 2022).

5.2 Resources for NER and RE

For applied domains, such as polymer science research, the availability of resources and datasets for identifying polymer names, relevant materials, and their relations is limited. As far as we know, there is no dataset that is solely focused on the recognition of named entities, extraction of relationships, and extraction of property data in scientific literature concerning polymers.

The dataset by Mysore et al. (2019) has 230 labeled synthesis procedures for inorganic synthesis. Yamaguchi et al. (2020) annotated a corpus of 1,000 abstracts about superconductive materials for NER purposes. O'Gorman et al. (2021) presented the largest NER dataset for materials science procedural text. Recently, Yang et al. (2022) released PcMSP, a corpus for entity and relation extraction from polycrystalline materials synthesis procedures. Shetty et al. (2023) annotated 750 polymer abstracts using their ontology for information extraction, which includes eight entity types but no entity relations.

6. Conclusion

In this study, we developed a novel ontology and a corpus that enables the simultaneous extraction of various entities and relations in the polymer science domain. PolyNERE presents several advantages, making it a reliable benchmark for related tasks. Our extensive experiments have achieved promising outcomes for both test abstracts and previously unseen paragraphs. Our future plans involve constructing a top-level ontology to enhance reuse and interoperability with related domains, and improving task performance across various settings to develop robust information extraction systems for the polymer domain.

Ethics Statement

The annotators did not receive payment, but they are all acknowledged as authors of this paper.

7. Bibliographical References

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

Chiu, J.P., & Nichols, E. (2015). Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4, 357-370.

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*.

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.

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF Models for Sequence Tagging. *ArXiv*, abs/1508.01991.

Kingma, D.P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980.

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*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*.

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2019). A Unified MRC Framework for Named Entity Recognition. *ArXiv*, abs/1910.11476.

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*.

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*.

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*.

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*.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y., & Yamazaki, M. (2011). PoLyInfo: Polymer Database for Polymeric Materials Design. *2011 International Conference on Emerging Intelligent Data and Web Technologies*, 22-29.

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.

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*.

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.

Shibuya, T., & Hovy, E.H. (2019). Nested Named

Entity Recognition via Second-best Sequence Learning and Decoding. *Transactions of the Association for Computational Linguistics*, 8, 605-620.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation. *Conference of the European Chapter of the Association for Computational Linguistics*.

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

Wang, J., Shou, L., Chen, K., & Chen, G. (2020). Pyramid: A Layered Model for Nested Named Entity Recognition. *Annual Meeting of the Association for Computational Linguistics*.

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*.

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.

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.

Yu, J., Bohnet, B., & Poesio, M. (2020). Named Entity Recognition as Dependency Parsing. *Annual Meeting of the Association for Computational Linguistics*.

Zhang, N., Chen, X., Xie, X., Deng, S., Tan, C., Chen, M., Huang, F., Si, L., & Chen, H. (2021). Document-level Relation Extraction as Semantic Segmentation. *ArXiv*, abs/2106.03618.

Zhou, W., Huang, K., Ma, T., & Huang, J. (2020). Document-Level Relation Extraction with Adaptive Thresholding and Localized Context Pooling. *AAAI Conference on Artificial Intelligence*.

A. Challenges of Polymer Science Domain

Polymer science presents unique complexities that distinguish it from many other domains:

- Diverse terminology: Polymers are characterized by multiple names and abbreviations, reflecting the diverse ways researchers refer to them based on structural features, synthesis methods, and other factors. Additionally, polymer science encompasses various forms, including homopolymers, copolymers, and polymer blends, further contributing to the complexity of the field. Furthermore, polymer science relies on specialized nomenclature systems, such as IUPAC (International Union of Pure and Applied Chemistry) recommendations for polymer names. This introduces an additional layer of complexity compared to more general naming conventions. As an illustration, consider the example from the PoLyInfo Database (Otsuka et al., 2011): while researchers commonly use the name "*nylon 5,11*" for a specific polymer, its corresponding IUPAC structure-based name is '*poly(iminopentane-1,5-diyliminoundecanedioyl)!poly(iminopentamethyleneiminoundecanedioyl)*', and its IUPAC source-based name is *'poly(pentamethylene undecanediamide)*'.

- Diverse properties: Polymers present a diverse range of properties, and the context used to describe these properties can vary in scientific papers. The inclusion of different units, scales, and representations introduces complexity both in developing a comprehensive ontology and in creating automatic information extraction systems.