Mention-Agnostic Information Extraction for Ontological Annotation of Biomedical Articles

Oumaima El Khettari[1,2*] **Noriki Nishida**[2*] **Shanshan Liu**[2] **Rumana Ferdous Munne**[2]
Yuki Yamagata[3,4] **Solen Quiniou**[1] **Samuel Chaffron**[1] **Yuji Matsumoto**[2]
[1]Nantes Université - LS2N [2]RIKEN AIP [3]RIKEN R-IH [4]RIKEN BRC
{oumaima.el-khettari, solen.quiniou, samuel.chaffron}@univ-nantes.fr
{noriki.nishida, shanshan.liu, rumanaferdous.munne,
yuki.yamagata, yuji.matsumoto}@riken.jp

Abstract

Biomedical information extraction is crucial for advancing research, enhancing healthcare, and discovering treatments by efficiently analyzing extensive data. Given the extensive amount of biomedical data available, automated information extraction methods are necessary due to manual extraction's labor-intensive, expertise-dependent, and costly nature. In this paper, we propose a novel two-stage system for information extraction where we annotate biomedical articles based on a specific ontology (HOIP). The major challenge is annotating relation between biomedical processes often not explicitly mentioned in text articles. Here, we first predict the candidate processes and then determine the relationships between these processes without relying on mentions. The experimental results show promising outcomes in mention-agnostic process identification using Large Language Models (LLMs). In relation classification, our proposed BERT-based models outperform LLMs significantly. The end-to-end evaluation results suggest the difficulty of this task and room for improvement in both process identification and relation classification.

1 Introduction

In the biomedical domain, unraveling the mechanisms underlying various diseases contributes significantly to their treatment and prevention. However, information about these mechanisms is often scattered across articles, presenting challenges. The challenges include the lack of clarity, the implicit nature of background knowledge, and the ad hoc use of vocabularies with variations in notation. Moreover, inherent biological complexity spans molecules, cells, and organs, with external factors such as viruses influencing infection mechanisms.

To address these challenges, organizing knowledge through ontologies is crucial as they provide a clear framework for consistently structuring entities and their relationships. In the Homeostasis Imbalance Process Ontology (HOIP), manual annotation has been employed to extract and structure knowledge about processes such as cellular senescence and COVID-19 infection mechanisms (Yamagata et al., 2021, 2024). Despite these systematic approaches, manual annotation faces significant challenges due to its high cost and time-consuming nature. These challenges highlight the need for more efficient and consistent (semi-)automated annotation approaches to improve the overall quality and usefulness of ontologies.

In this paper, we propose an application of Natural Language Processing (NLP) as a promising solution. Specifically, assuming automatic annotation of the HOIP ontology as our ultimate goal, we propose a two-stage information extraction (IE) system. Figure 1 shows an overview of our two-stage system. Given an input passage[1], the first stage, *Process Identification*, identifies process entities that are described in the passage or can be inferred using the domain knowledge.[2] The entities are represented as unique IDs in the ontology. The entities are then passed to the second stage, *Document-level Relation Extraction (DocRE)* (Christopoulou et al., 2019; Zhou et al., 2021; Xiao et al., 2022; Zhang et al., 2021; Li et al., 2023), to classify entity pairs into a pre-defined set of interrelations. The system output is represented as a set of triples: {(head entity ID, relation, tail entity ID)}. We develop and evaluate different approaches including supervised models based on BERT (Devlin et al., 2019) and

[*]Equal contribution.

[1]In this paper, we use the word "passage" instead of "document" or "paragraph" to describe the input in our task, because the text describing biomedical processes is not necessarily a complete text like an entire paragraph or document.

[2]Process Identification is similar to Entity Disambiguation (ED), but differs as discussed in the following paragraph. Given an input passage, ED aims to identify entities (IDs) for each given mention, whereas Process Identification aims to identify entities (IDs) without the availability of mentions.

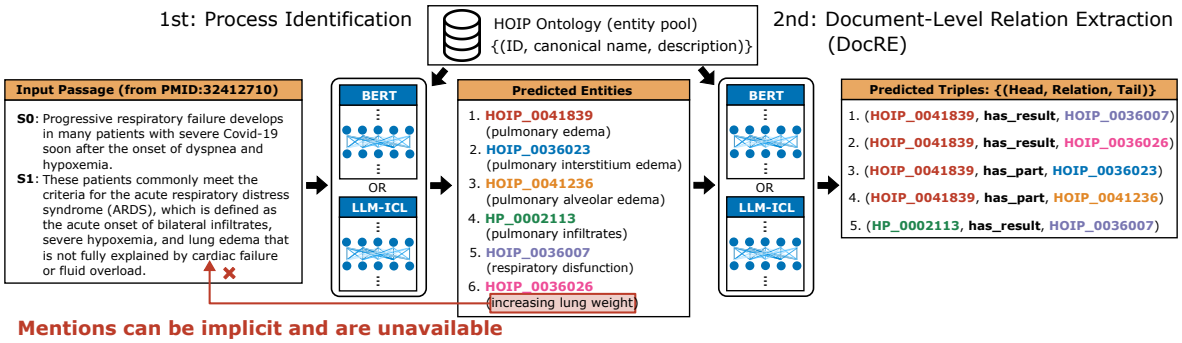

Figure 1: An overview of our mention-agnostic two-stage information extraction system with a real example in our HOIP dataset. Given an input passage, the first stage identifies process entities described in the passage or inferable based on the domain knowledge. The predicted entities are then passed to the second stage to identify relations between them. Please note that our system does not rely on mentions, enabling extraction of structured knowledge about entities and relations described implicitly in the passage.

generative methods based on Large Language Models (LLMs) and In-Context Learning (ICL) (Brown et al., 2020; Chowdhery et al., 2022; Wadhwa et al., 2023; Ozyurt et al., 2024) for both process identification and DocRE. The HOIP dataset, a novel manually annotated dataset built based on the HOIP ontology for biomedical IE system development, will be available to the public.

Traditional IE studies (Yu et al., 2020; Wu et al., 2020; Zhou et al., 2021) assume that an entity can appear multiple times in a passage **explicitly** (such textual instances are called *mentions*), and derive entity features from these mentions. Mentions are strong indicators in IE, since they directly indicate how entities are described in a text. However, in real-world scenarios including our HOIP dataset, an entity sometimes appears only **implicitly**. With no availability of mentions, it is not obvious how to induce useful entity features from a passage. This paper proposes multiple approaches that do not require explicit mentions.

Our contributions and findings are summarized as follows:

- We release the HOIP dataset, to facilitate the development and bench-marking of IE models for the real-world ontology.
- We develop a mention-agnostic two-stage IE system, which enables to extract structured knowledge described implicitly in text. BERT-based supervised models and LLM-based models are presented for both process identification and DocRE.
- Experimental results in process identification suggest that generative models are valuable for low-resource in-domain corpora like the HOIP dataset.
- DocRE results suggest that, although mentions are strong indicators, the proposed BERT-based models outperform LLMs and achieve F1 scores of around 56-59 points even without mention hints.
- Evaluation results on the end-to-end system reveal that improvements in both process identification and DocRE are crucial in the current stage.
- The HOIP dataset and the source codes are available: `https://github.com/norikinishida/hoip-dataset` (dataset), `https://github.com/sl-633/bio-process-identifier` (process identification), `https://github.com/norikinishida/kapipe` (DocRE).

2 HOIP Dataset

Our ultimate goal is to update and improve ontologies by (semi-)automatically extracting entities and their interrelations from articles. As a testbed ontology, we choose the Homeostasis Imbalance Process Ontology (HOIP) (Yamagata et al., 2021, 2024), which focuses on understanding the COVID-19 infectious mechanism (courses).[3] To facilitate the development of NLP systems and benchmark the task, we construct and release a new dataset named the HOIP dataset based on the HOIP ontology. The dataset includes passages extracted from PubMed articles describing biomedical processes in the context of COVID-19 infectious courses. Each passage is a brief portion of a PubMed article that describes at least two specific processes. The processes are manually annotated as a set of triples, i.e., {(head entity, relation, tail entity)}. Figure 1 shows a real example in the dataset.

[3]For the details of the ontology see Appendix A.

	Train	Dev	Test
# passages	255	35	37
# entities	1988	143	211
# triples	1848	137	177
Avg. words per passage	75.5	70.4	61.8
Avg. entities per passage	7.8	4.1	5.7
Avg. triples per passage	7.2	3.9	4.8

Table 1: Dataset statistics for the HOIP dataset.

2.1 Data Collection and Enhancement

We first stored the HOIP ontology files in an RDF store using Apache Jena Fuseki[4], and constructed a SPARQL endpoint. We used SPARQL queries to retrieve the information required for the dataset. The results were then converted to CSV. To optimize the dataset for machine annotation and enhance its clarity, we made several adjustments based on the hierarchical structure of the HOIP ontology. Originally, some process entities included *course* information, such as "blood vessel damage *in severe COVID-19*" (the course is in italics), indicating a specific context. We removed these course information from the entities to optimize for machine annotation. Additionally, processes with too fine granularity were deemed unsuitable for machine annotation predictions. Therefore, we prioritized processes that are generalized using superclasses of each process, assigning Gene Ontology (GO) terms (Ashburner et al., 2000). This approach ensures that the annotations are practical for applications for reusability.

2.2 Dataset Organization

In the CSV file generated by the above procedure, each record corresponds to one triple. We combined the triples associated with the same passage (string) and PubMed ID into the same group, and this group was considered to be a single example in the final dataset. We found that there were textual overlaps across the passages. Thus, if a passage d_{src} was textually contained in another passage d_{dst} and both passages are associated with the same PubMed ID, the triples T_{src} for d_{src} were merged into the triples T_{dst} for d_{dst}. Finally, we split the entire dataset into training, development, and test sets, ensuring that passages extracted from the same article were not scattered across different splits. The dataset statistics are shown in Table 1.

[4]https://jena.apache.org/documentation/fuseki2/

3 Methods for Process Identification

In the HOIP dataset, a biological process entity is annotated depending on whether it is mentioned (explicitly or implicitly) in the passage, without specifying the corresponding phrase of the entity in the passage. This makes the dataset more closely match the real-world scenario, but also brings challenges to the automatic process identification – directly employing Named Entity Recognition methods that require the correspondence between a explicit mention (entity text and offsets) and an input text for model training is no longer an option. To address this task, we propose approaches to identify biological processes without prior recognition of mentions that can be matched to terminological expressions of entities in the HOIP ontology. Two distinct approaches are developed: BERT-based supervised methods and LLM-based In-Context Learning (ICL) methods.

3.1 BERT-based Supervised Approach

Considering that 360 unique process names are encompassed in the HOIP dataset, the task of process identification can be approached as a multi-class and multi-label classification problem. For simplification purposes, we convert the task into a binary format, framing it in the following manner: Let D be the set of passages and A be the set of annotated process names. The input sequence is constructed for each passage from D as follows:

$$[\texttt{CLS}]\ \text{passage}\ [\texttt{SEP}]\ \text{name}\ [\texttt{SEP}]$$

where name denotes a process name $a_i \in A$. Then, the task is a binary classification task whether the passage involves the process or not.

3.2 LLM-based ICL Approach

Taking into account the unique characteristics of the dataset and the rapid advancements in the capabilities of LLMs to produce coherent texts in low resource settings (Wang et al., 2023), LLMs are utilized in this study to generate HOIP processes for each passage. We aim to evaluate the model's performance in low-resource settings characterized by imbalanced data in a specialized domain, and assess the model's generative capability in producing HOIP ontology terms.

- Zero-shot setting: The model is prompted to list the biological processes present in the text. Following a prompt format being demonstrated effective in many studies (Mishra et al., 2022;

Sclar et al., 2023), the prompt includes task instruction, constraints on the output, the input text. An example of the prompt is shown in Table 8.

- Few-shot setting: Following the previously mentioned prompt format, two few-shot strategies are employed through adding demonstrations: the first involves selecting randomly three examples from the development set, while the second is selecting examples based on semantic closeness of process names.

4 Methods for Document-level RE

Given an input passage d, a set of entities for the passage $\{e_1, \cdots, e_K\}$, and a pre-defined set of relations $\mathcal{R}$, document-level relation extraction (DocRE) (Christopoulou et al., 2019; Zhou et al., 2021; Xiao et al., 2022; Zhang et al., 2021) aims to predict relations from $\mathcal{R} \cup \{\text{NA}\}$ for entity pairs (e_i, e_j) $(i, j \in [1, K]; i \neq j)$, where e_i and e_j denote head and tail entities respectively, and the NA class indicates that the entity pair has no relation.

4.1 QA-Style DocRE Model

Our first approach is to perform DocRE as a Question Answering (QA) task. We first generate questions for each possible triple. The question and the input passage are concatenated and passed to a pre-trained language model for answering.

Question Generation. We first enumerate all possible entity pairs $\{(e_i, e_j)\}_{i,j \in [1,K]; i \neq j}$, and then apply pre-defined template functions $\{\mathcal{T}_r\}_{r \in \mathcal{R}}$ to the entity pairs to obtain questions for *each* possible triple (e_i, r, e_j): $q_{(e_i,r,e_j)} = \mathcal{T}_r(e_i, e_j)$. Table 7 shows examples for the pre-defined templates. The input x to our QA model is as follows:

$$\texttt{[CLS]}\ \text{question}\ \texttt{[SEP]}\ \text{passage}\ \texttt{[SEP]}$$

where question and passage are the word-pieces tokens of $q_{(e_i,r,e_j)}$ and d, respectively. An example for the input is "[CLS] does immunoglobulin production result in immunoglobulin mediated immune response ? [SEP] within 19 days after symptom onset , 100 % ... [SEP]".

Answer Classification. Then, we feed the input sequence x into a BERT-based encoder (Devlin et al., 2019; Beltagy et al., 2019) to obtain the contextual embeddings: $\{\boldsymbol{h}_w\}_{w=1}^{N^x_{\text{tok}}} = \text{Encoder}(x)$, where N^x_{tok} is the number of tokens in x. We concatenate the output of the last layer for the `[CLS]` token and the average-pooling embedding to obtain the passage embedding: $\tilde{\boldsymbol{h}} = \boldsymbol{h}_1 \oplus \frac{1}{N^x_{\text{tok}}} \sum_{w=1}^{N^x_{\text{tok}}} \boldsymbol{h}_w$, where $\oplus$ represents the concatenation of vectors. Then, we apply a two-layer feed-forward network and sigmoid activation to the passage embedding to calculate the probability of answer "Yes".

Loss Function. The network is trained using a binary cross entropy loss to maximize the probability for the correct triples.

4.2 Mention-Agnostic ATLOP (MA-ATLOP)

Our first approach requires solving QAs for all the possible triples. Since the number of possible triples is increased by $O(K^2)$ for the number of entities K, this is not efficient. Our second approach is to make predictions over all possible triples in a single forward pass. We extend a traditional and popular DocRE method, ATLOP (Zhou et al., 2021), so as not to rely on explicit mentions. We call this method *Mention-Agnostic ATLOP*, or MA-ATLOP shortly.

Entity Encoding. We use a BERT-based encoder to encode each entity e_i and passage d jointly into a dense vector that takes into account how the entity e_i is described in the passage d. Specifically, we first retrieve the canonical names $\{n_i\}_{i=1}^K$ and the descriptions $\{s_i\}_{i=1}^K$ for the given entities from the ontology using the entity IDs as query. Then, for each entity e_i, we construct input x_i as follows:

$$\texttt{[CLS]}\ \text{name} : \text{description}\ \texttt{[SEP]}\ \text{passage}\ \texttt{[SEP]}$$

where name, description, and passage are the word-pieces tokens of n_i, s_i, and d, respectively. We apply the encoder to each input x_i independently to obtain contextual embeddings: $\{\boldsymbol{h}_{i,w}\}_{w=1}^{N^{x_i}_{\text{tok}}} = \text{Encoder}(x_i)$. We take the embedding of the `[CLS]` token as the entity embedding, i.e., $\boldsymbol{e}_i = \boldsymbol{h}_{i,1}$.

Relation Classification. After obtaining the entity embeddings $\{\boldsymbol{e}_i\}_{i=1}^K$, we apply two separate FFNNs and tanh activation to map them to different representations for the head/tail entities of triples. Then, we apply a group bilinear classifier (Zheng et al., 2019; Tang et al., 2020) to the head/tail representations of an entity pair (e_i, e_j). Specifically, we divide both head/tail representations into G contiguous groups and then apply bilinear to each group. They are then summed up to calculate the score for relation $r \in \mathcal{R} \cup \{\text{TH}\}$. Refer to Zhou et al. (2021) for the detail of group bilinear. We follow ATLOP and employ the adaptive-thresholding

class TH. The relations scored higher than the TH class are regarded as positive. If no such relation exists, the NA class is assigned to the entity pair.

Loss Function. We use the adaptive-thresholding loss proposed in ATLOP to push the scores of correct/incorrect relations to be higher/lower than the TH class.

Negative Entity Sampling (NES). In an experimental setting, we assume experts correctly annotate entities. However, in real-world situations, entities are automatically annotated by the systems. Thus, it often happens that entities not described in the passage are included in the given entity list $\{e_i\}_{i=1}^{K}$. Our DocRE system must be robust to such noisy (false-positive) entities. Therefore, we propose Negative Entity Sampling (NES), where we sample additional negative entities randomly and add them to the given entities $\{e_i\}_{i=1}^{K_{\text{pos}}}$ during training. We sample negative entities from all entities in the ontology. Given the number of positive entities K_{pos} and a hyperparameter $\rho > 0$, we define the number of sampled negative entities as $K_{\text{neg}} = \text{round}(\rho \times K_{\text{pos}})$, where round is the rounding function. For instance, for $K_{\text{pos}} = 10$ and $\rho = 0.5$, K_{neg} is 5. We add a linear layer to the network to classify whether the entity is described in the passage: $y_i^{\text{ent}} = \sigma(\text{FFNN}_{\text{ent}}(\boldsymbol{e}_i))$. We use a binary cross entropy as an auxiliary loss to maximize y_i^{ent} for positive entities.

4.3 LLM and In-Context Learning for DocRE

To investigate the effectiveness of LLMs with In-Context Learning (ICL) (Brown et al., 2020; Chowdhery et al., 2022; Wadhwa et al., 2023; Ozyurt et al., 2024) in our task, we compare the LLM-ICL results with the above BERT-based models. Table 9 in Appendix C shows a prompt example we used in our experiments. Specifically, we instruct an LLM to generate a bulleted list of triples for the given passage and the entity list.[5] Each entity is represented in the form "* <ID> : <NAME>" and the entity list is presented as a bulleted list. In the prompt, we also use 3 examples randomly sampled from the training set as the few-shot demonstrations. The same demonstrations are used for all test passages. From each bulleted line generated, we extract the head entity ID e_i, the relation label r, and the tail entity ID e_j using regular expressions. If the extracted entity IDs (e_i, e_j) and the relation label (r) cannot be found in the given entity list $\{e_1, \cdots, e_K\}$ and the possible relation classes $\mathcal{R}$, the bulleted line is ignored. We also remove duplicated triples. The resulting triples are then compared with the gold triples for evaluation.

[5] In our preliminary experiments, we also tried to generate JSON directly by Llama2 13B; however, generating JSON yielded lower DocRE scores consistently than generating text.

5 Experiments on Process Identification

Binary classification. In the supervised task formulation, each passage is associated with all 360 process names, with binary labels assigned based on the presence of them in the annotations. Numerous negative samples are constructed for each passage. Consequently, we incorporate negative sampling using various ratios of negative to positive samples. The classification task involves fine-tuning BERT-based models – BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), and PubMedBERT (Gu et al., 2021) on the training set. For hyper-parameter details see Table 11.

Generative experiments. Two instruction types are utilized: One prompts the model to list all biological processes in the text, while the other instructs it to generate pairs of processes having a relation. This distinction stems from annotation being conducted at the relation level, where only processes involved in relations are annotated. Consequently, other processes may exist in the text but aren't annotated. We employ Llama2 13B (Touvron et al., 2023) and Llama3 8B (AI@Meta, 2024) on the test set to ensure comparability with the supervised method's results.

5.1 Clustering-based Demonstration Selection

In the zero-shot setting, process names differ significantly from the provided annotations. In the few-shot scenario, performance is highly sensitive to the chosen demonstrations (Li et al., 2022; Lu et al., 2022; Zhang et al., 2023). To enhance few-shot performance, we introduce a retrieval module based on semantic similarity to cluster the most relevant examples from the annotated processes in the development set.

To achieve this, we use the development set to create 10 clusters with K-means clustering (Lloyd, 1982; MacQueen et al., 1967), based on the annotated process list for each passage. Each list is encoded into a vector using BERT by averaging the last hidden state of the [CLS] token for all labels. Each passage in the test set is assigned a cluster given the last hidden state of the [CLS] token of the

Positive-Negative Ratio	BioBERT			PubMedBERT			SciBERT		
	P	R	F1	P	R	F1	P	R	F1
1:8	21.7	27.9	**24.4**	22.2	42.6	**29.2**	21.3	18.9	20.1
1:4	15.8	45.0	23.4	18.4	60.6	28.2	17.6	44.5	**25.3**
1:1	10.6	64.9	18.3	12.0	47.3	19.2	9.99	71.5	17.5

Table 2: Results of the BERT-based supervised approach on process identification. Precision (P), Recall (R), and F1 scores on the test set of the HOIP dataset are reported. Values in bold represent the best F1 score for each model.

Method	Top-K	In-Ontology Matching			In-Dataset Matching		
		P	R	F1	P	R	F1
Llama2 13B	1	10.7	22.3	14.5	11.2	28.9	16.2
	3	10.0	30.8	15.1	6.7	45.5	11.7
	5	8.3	32.7	13.3	5.4	57.3	9.9
	10	7.0	38.9	11.9	4.1	74.9	7.8
Llama3 8B	1	43.1	11.8	18.6	34.3	28.4	31.1
	3	39.5	16.1	22.9	19.5	43.6	26.9
	5	33.9	19.0	24.3	15.6	55.9	24.4
	10	29.4	27.5	28.4	9.5	62.1	16.5

Table 3: Results of the LLM-based in-context learning approach on process identification. For In-Ontology Matching, we used all entities found in the HOIP ontology as the candidate entities for matching. Matched entities that are not present in the HOIP dataset are ignored. For In-Dataset Matching, we used entities only found in the HOIP dataset as the candidate entities for matching.

passage to the clustering model. During generation, examples from the assigned cluster are included in the prompt. See Table 10 for details.

5.2 Evaluation Methods

We evaluate three primary aspects of generative process identification. The initial aspect involves the Direct Output assessment, where we directly evaluate the output by comparing it to the annotation. This process aims to weight the model's ability to generate processes formulated within the knowledge base framework. As for the second aspect – assessing the system's capability to automatically populate the target knowledge base, we include an ontology alignment-based evaluation which is presented in three steps:

1. Computing embeddings: Let $E_{\text{generated}}$ be the set of embeddings of the generated processes and E_{ontology} be the set of embeddings of ontology terms calculated by SapBERT (Liu et al., 2021).
2. Selecting top k generated processes: For each x_i from $E_{\text{generated}}$, calculate $\text{sim}(x_i, y_j)$ for all y_j from E_{ontology}. Then, select the k elements with the highest cosine similarity.
3. Computing precision, recall and F1 scores between the list of annotated processes and the flattened list of top k generated processes at a passage level.

The last aspect of evaluation mirrors the second one, with the distinction being that instead of matching with the entire ontology, only the process names used in the dataset are taken into account. This approach is grounded on the assumption that the terms utilized in the dataset annotation are the most prevalent.

5.3 Results and Discussion

Direct output assessment. We compared annotation labels with generated process names using zero-shot, regular few-shot, and ICL few-shot settings. The ICL few-shot method achieved the most exact matches, with 30 compared to 2 for regular few-shot and none for zero-shot, underlining the importance of better selected demonstrations, as in Min et al. (2022). Thus, evaluation using the HOIP ontology and dataset matching will be based on the ICL few-shot setting outputs.

In-Dataset Matching. We report in Table 2 the results of the fine-tuned BERT-based models. The negative ratio significantly influences the overall performances of the models. Results indicate that with fewer negative samples, models are more likely to identify true positives but at the cost of also misclassifying more false positives. This is likely due to semantic similarity among the inputs.

PubMedBERT achieves the highest F1 score, under the optimal negative ratio of 8.

Furthermore, we compare the results of the supervised approach with the top-1 results of Llama2 and Llama3 present in Table 3. Across all negative ratios, the F1 score of BERT-based models exceeds the results of Llama2, even under the low-resource setting. However, this trend changes with Llama3, which outperforms the PubMedBERT result by nearly 2 points. Comparing Llama2 and Llama3 reveals that Llama3 is more effective at generating well-tailored process names, resulting in higher precision. Llama3 generates fewer, but better-quality candidates, enhancing task performance. Llama2 improves with more candidates, increasing chances of correct matches, but accuracy still depends significantly on the quality of these generated candidates.

In-Ontology Matching. Following the same tendency in the In-Dataset setting, matching with better-generated process names proves to be more effective overall. Since the goal of this matching is to automatically populate an ontology and DocRE is the next step in the pipeline, concentrating on finding the correct process names is crucial. This focus will aid the DocRE step in serving as a filtering mechanism, ensuring more accurate and relevant candidate triplets to be added to the ontology.

6 Experiments on Document-level RE

We evaluate our systems on the HOIP dataset and CDR dataset (Li et al., 2016). CDR consists of 1,500 abstracts from PubMed, manually annotated with Chemical or Disease entities and Chemical-Induce-Disease relations between them. Since entity IDs in CDR are MeSH unique IDs (e.g., D006493), and the HOIP ontology, while highly specialized for annotating processes related to COVID-19, does not provide the coverage for general chemical compounds and disease terms as the MeSH controlled vocabulary, we used MeSH instead of the HOIP ontology for CDR. We use precision, recall, and F1 metrics, and report the scores averaged independently over 3 trials with different random seeds. A triple (e_i, r, e_j) is considered correct when the head entity ID (e_i), the tail entity ID (e_j), and the relation label (r) are all predicted correctly. We used greedy decoding in the LLM-ICL methods, the results of which do not depend on the seed differences. Table 12 shows the hyperparameters for our DocRE models.

Method	P	R	F1
ATLOP (all mentions)	**64.61**	**75.92**	**69.74**
Llama3 8B (all mentions)	42.26	48.69	45.25
QA-Model (first mention)	56.40	67.39	61.36
MA-ATLOP (first mention)	57.54	**68.11**	**62.34**
MA-ATLOP (first mention) + NES	**57.55**	67.95	62.31
Llama3 8B (first mention)	43.62	49.34	46.30
QA-Model	53.37	64.01	58.12
MA-ATLOP	53.72	65.92	59.18
MA-ATLOP +NES	**54.03**	**66.20**	**59.50**
Llama3 8B	44.75	51.97	48.09

Table 4: DocRE results on the CDR test set. All metrics are averaged over 3 trials. "all mentions" (or "first mention") indicates that the models use all mentions (or the first-appearing mention) as the entity names instead of the canonical name retrieved from the MeSH ontology. The best scores are in bold for each block.

6.1 Experiments on the CDR Dataset

To investigate the importance of mentions in DocRE and how well our models can identify relations without relying on mentions, we first evaluate the models on CDR. We also evaluate a variant of each of our models, which uses the first-appearing annotated mention as the entity name rather than the canonical name retrieved from the MeSH ontology. Although this variant still does not use mention spans, we expect this variant to recognize more easily how the entity is described in the passage than the original model, because the entity names appear at least once in the passage.

Table 4 shows the results. ATLOP exploits mention spans as the direct hints for entity encoding and achieves an F1 score of 69.7. In contrast, our best mention-agnostic model, i.e., MA-ATLOP (+ first mention), achieved an F1 score of 62.3, lower than the ATLOP score by 7.4 points. When there were no mention hints at all, MA-ATLOP and QA-Model yielded F1 scores of 59.2 and 58.1, respectively. These results suggest that our models can identify triples more accurately than expected even without mention hints; however, mentions are still crucial in this task. Also, the BERT-based supervised models outperformed the LLM counterparts. MA-ATLOP outperformed QA-Model consistently. Considering that MA-ATLOP also has higher computation efficiency than QA-Model, MA-ATLOP is more suitable for real-world applications. By employing Negative Entity Sampling (NES), when no mention is available, MA-ATLOP improved all metrics slightly, suggesting the effectiveness

Method	Entity	P	R	F1
QA-Model	gold	51.5	**63.1**	56.7
MA-ATLOP	gold	67.2	52.6	**58.9**
MA-ATLOP + NES	gold	**71.2**	48.6	57.7
Llama3 8B	gold	18.5	16.7	17.6
Upper-bound	pred.	100.0	26.8	42.3
MA-ATLOP	pred.	7.7	14.9	10.2

Table 5: DocRE results on the HOIP test set. The upper and lower blocks show the results when using the ground-truth entities or predicted entities, respectively. The predicted entities are provided by Llama3 8B.

of NES. For the "first-mention" setting, NES did not improve the performance, probably due to the discrepancy between the entity-name style between positive entities (mention) and negative entities (ontology-based name).

6.2 Experiments on the HOIP Dataset

The upper block in Table 5 shows the results on the HOIP dataset when using the ground-truth entities. We evaluate the models that do not require mention hints on this dataset. The BERT-based models achieved much higher F1 scores (56.5-59.0) than LLM (17.6). MA-ATLOP also outperformed QA-Model by 2.2 points in F1. These results were consistent with the results on CDR, demonstrating the effectiveness of MA-ATLOP in terms of both accuracy and computational efficiency in this task. Negative Entity Sampling improved the precision by 4 points, but decreased the recall. These results suggest that, while NES enhances the filtering capability of MA-ATLOP, NES also has the effect of making the model reluctant about positive predictions, and it would be necessary to develop techniques to avoid such biases.

The above experiments assume that the entities are fully and correctly annotated. This setup is appropriate for a clean measurement of the DocRE system's performance itself. However, in reality, entities can be predicted automatically. To evaluate the whole system's performance in the real-world situations, we evaluate our best DocRE model (MA-ATLOP) on the HOIP dataset with entities predicted by Llama3 (8B).

The lower block in Table 5 shows the results. We first calculated the upper-bound scores for the predicted entities. Specifically, we created a subset of gold triples that can be created based on the predicted entities. Precision, recall, and F1 scores are 100.0, 26.8, and 42.3, respectively. The precision does not depend on the quality of predicted entities. The lower recall suggests that there is much room for improvement in DocRE as the process identification recall improves. MA-ATLOP yielded 7.7, 14.9, and 10.2 scores for precision, recall, and F1, respectively. Compared to the much higher precision (67.2) in the gold-entity setup (Table 5), the results suggests that the current model struggles to filter out noisy triples with irrelevant entities. In summary, both improvements in recall (coverage) and precision (low-noisiness) in process identification and DocRE are needed in the current situation, suggesting the difficulty of this task.

7 Case Study

We performed case study to analyze the system outputs qualitatively. We used the best Llama3 (8B) and MA-ATLOP models for process identification and DocRE, respectively.

Table 6 shows an example with true-positive and false-negative entities and triples. Additional example can be found in Appendix F. For ease of understanding, entities are shown by names, not by IDs. We can observe that entities that are almost explicit in the passages, such as "pyroptosis" (*Pyroptosis* in the passage), and "pore formation in membrane of other organism" (*Formation of pores*), were accurately extracted. Triples that are almost explicit based on the context, such as ("pyroptosis" has part, "pore formation in membrane of other organism") and ("pyroptosis", has result, "release of DAMP molecules by cell rupture"), were correctly identified by the DocRE system. In contrast, implicit (or knowledge-requiring) entities and triples, such as "binding of pattern recognition receptor to DAMPs", were not identified by the systems. The entity is derived from the interpretation that *these molecules recruit more immune cells*, which requires background knowledge of immunology: DAMP molecules must bind to receptors recognized by immune cells to recruit immune cells.

The quality evaluation reveals several insights: (1) Detailing causal relationships in elucidating disease mechanisms often necessitates background knowledge not explicitly mentioned in articles. This background knowledge is sometimes added to intermediate causal entities by manual annotation. The BERT-based supervised models and LLMs have difficulty in obtaining such background knowledge and understanding the task from limited

ID72. Passage:
Pyroptosis is a highly inflammatory form of lytic programmed cell death that occurs most frequently upon infection with intracellular pathogens and is likely to form part of the antimicrobial response. Pyroptosis can take place in immune cells and is also reported to occur in keratinocytes and some epithelial cells. Formation of pores causes cell membrane rupture and release of cytokines, as well as various damage-associated molecular pattern (DAMP) molecules such as HMGB-1, ATP and DNA, out of the cell. These molecules recruit more immune cells and further perpetuate the inflammatory cascade in the tissue.
True-Positive Entities:
* pore formation in membrane of other organism
* pyroptosis
* release of DAMP molecules by cell rupture
False-Negative Entities:
* binding of pattern recognition receptor to DAMPs
True-Positive Triples:
* (pyroptosis, has part, pore formation in membrane of other organism)
* (pyroptosis, has result, release of DAMP molecules by cell rupture)
False-Negative Triples:
* (release of DAMP molecules by cell rupture, has result, binding of pattern recognition receptor to DAMPs)

Table 6: A case study on our process identification and DocRE models. Phrases referred in Section 7 are underlined.

supervision (labeled data or demonstrations). (2) NLP systems could be complementary to manual annotations. Manual annotation often focuses on the causality of a particular process in one literature and gives priority to further causes and consequences in other literature. Therefore, other processes and causal relationships in the same passage may not be extracted. It is also possible that processes that missed identification due to simple errors due to annotation fatigue are also in the false positive. In such manual annotation issues, NLP analysis could make a significant contribution to the identification of processes.

8 Related Work

Gene Ontology Causal Activity Modeling (GO-CAM) defines molecular-level causal relationships (Thomas et al., 2019); however, it lacks granularity and context for COVID-19 infection. Our HOIP dataset is based on the HOIP ontology (Yamagata et al., 2021, 2024), which organizes knowledge about biomedical processes in the context of COVID-19 infectious courses and thus essential for analyzing SARS-CoV-2 infection and progression.

In knowledge acquisition, entities are typically identified by Named Entity Recognition (NER) and Entity Disambiguation (ED). The task of NER is to identify mentions in the given text that represent one of the pre-defined types (e.g., Chemical, Disease) (Yu et al., 2020; Zhu and Li, 2022; Ye et al., 2022). The entity mentions are then passed to ED to link them to the knowledge-base concept IDs that the mentions refer to best (Kolitsas et al., 2018; Wu et al., 2020; Cao et al., 2021; Yamada et al., 2022). These tasks commonly assume that entities are explicitly described in text. In reality, however, entities are not necessarily explicitly described. In this work, we explore mention-agnostic methods for process identification.

The most widely used approach to DocRE is to model entities by a pre-trained Transformer and perform pairwise relation classification. Christopoulou et al. (2019) proposed to model entity dependencies via graphs with nodes of various granularities. Zhou et al. (2021) proposed ATLOP, which models entity-pair contexts for pairwise relation classification. Xiao et al. (2022) introduced evidence modeling for improving ATLOP. Zhang et al. (2021) used U-Net architecture for modeling entity dependencies. These methods commonly rely on mentions and often insert special mention-boundary markers into text to indicate the mention locations to the Transformer. However, Li et al. (2023) showed that these methods are too sensitive to the accuracy of mentions and it is unrealistic to expect perfect mentions in the real-world scenario. In contrast, we propose mention-agnostic DocRE methods and investigate how well the mention-agnostic models can identify relations.

9 Conclusion

To assist ontology-based biological knowledge annotation, this work proposes a new dataset and practicable entity- and relation-level biomedical information extraction methods. We will continue to promote relevant research of semi-automatic annotation and advance practical applications.

Limitations

Despite demonstrating promising outcomes in mention-agnostic process identification and DocRE, our methodology does face limitations. First, our two-stage IE system consists of a cascade of process identification and DocRE, which inevitably suffers from error propagation. The experimental results in the pipeline setting suggest that the DocRE performance is significantly vulnerable to the accuracy of predicted entities. Moreover, the process identification model and the DocRE model are disconnected and cannot interact with each other. Second, our methods have only been evaluated in the domain of the HOIP ontology, and the accuracy in other biomedical domains and ontologies remains unknown. Third, our methodology has not been fully evaluated by domain experts. Although an expert analysis is performed, the analysis is based primarily on just two examples. A more thorough and detailed analysis by specialists is needed. Tackling these limitations remains an intriguing avenue for future research.

Acknowledgments

We would like to thank the anonymous reviewers for their thoughtful and insightful comments, which we found very helpful in improving the paper. This work was supported by JSPS KAKENHI Grant Numbers JP22K17959, 21K17815, and JP22H05015, and ANR AIBy4 project (ANR-20-THIA-0011), Nantes University.

References

AI@Meta. 2024. Llama 3 model card.

M. Ashburner et al. 2000. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25:25–29.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. *Preprint*, arXiv:2010.00904.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4925–4936.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database : the journal of biological databases and curation*.

Jing Li, Yequan Wang, Shuai Zhang, and Min Zhang. 2023. Rethinking document-level relation extraction: A reality check. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5715–5730, Toronto, Canada. Association for Computational Linguistics.

Tianyi Li, Wenyu Huang, Nikos Papasarantopoulos, Pavlos Vougiouklis, and Jeff Z. Pan. 2022. Task-specific pre-training and prompt decomposition for knowledge graph population with language models. *ArXiv*, abs/2208.12539.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *Preprint*, arXiv:2202.12837.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Yilmazcan Ozyurt, Stefan Feuerriegel, and Ce Zhang. 2024. Document-level in-context few-shot relation extraction via pre-trained language models. *Preprint*, arXiv:2310.11085.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Yun Tang, Jing Huang, Guangtao Wang, Xiaodong He, and Bowen Zhou. 2020. Orthogonal relation transforms with graph context modeling for knowledge graph embedding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2713–2722, Online. Association for Computational Linguistics.

P.D. Thomas, D.P. Hill, H. Mi, et al. 2019. Gene ontology causal activity modeling (go-cam) moves beyond go annotations to structured descriptions of biological functions and systems. *Nat Genet*, 51:1429–1433.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Qinyong Wang, Zhenxiang Gao, and Rong Xu. 2023. Exploring the in-context learning ability of large language model for biomedical concept linking. *Preprint*, arXiv:2307.01137.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. SAIS: Supervising and augmenting intermediate steps for document-level relation extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2395–2409, Seattle, United States. Association for Computational Linguistics.

Ikuya Yamada, Koki Washio, Hiroyuki Shindo, and Yuji Matsumoto. 2022. Global entity disambiguation with BERT. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3264–3271, Seattle, United States. Association for Computational Linguistics.

Yuki Yamagata, Tsubasa Fukuyama, Shuichi Onami, and Hiroshi Masuya. 2024. Prototyping an ontological framework for cellular senescence mechanisms: A homeostasis imbalance perspective. *Sci Data*, 11:485.

Yuki Yamagata, T. Kushida, Shuichi Onami, and Hiroshi Masuya. 2021. Ontology development for building a knowledge base in the life science and structuring knowledge for elucidating the covid-19 mechanism. In *Proceedings of the Annual Conference of JSAI*, pages 3H1GS3d01–03H1GS03d01.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *Preprint*, arXiv:2301.07069.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. 2019. Learning deep bilinear transformation for fine-grained image representation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

A HoIP Ontology

Understanding the HOIP ontology may be helpful for understanding our HOIP dataset and the task. In this section, we describe the features of the HOIP ontology.

The HOIP ontology is annotated based on COVID-19 related articles in PubMed using Protégé 5.5.0[6] and the Web Ontology Language (OWL). The COVID-19 infectious processes are manually annotated. Passages corresponding to the annotated terms are also provided. Article identifiers (e.g., PubMed ID (PMID: 25301932), DOI) are also provided using the database cross-reference annotation property.

The processes in HOIP consist of a hierarchy. The infectious processes described in the articles and the superclass of each process using Gene Ontology are annotated.

The relationships between processes are annotated using object properties, Causal relationships between processes are primarily annotated using the 'has result' relationship. Furthermore, subprocesses of a process are identified using the 'has part' relation.

HOIP defines a "COVID-19 infectious course" as a sequence of the abovementioned processes to describe infectious mechanisms. These courses are organized into an is-a (subclass of) hierarchy by severity, ranging from mild to severe. Notably, the "COVID-19 severe course" includes a subclass associated with acute respiratory distress syndrome (ARDS). These COVID-19-specific processes are used as our primary dataset for this study.

B Question Templates

Table 7 shows the question templates $\mathcal{T}_r$ ($r \in \mathcal{R}$) used for QA-Model. In the table, <HEAD> and <TAIL> are replaced by the head and tail entity names, respectively. The entity names are retrieved from the ontology using the entity IDs as query. We

[6] https://protege.stanford.edu

Dataset	Relation	Question Template
CDR	CID	*Does <HEAD> induce <TAIL> ?*
HOIP	has result	Does <HEAD> result in <TAIL> ?
	has part	*Does <HEAD> involve <TAIL> ?*
	has molecular reaction	*Does <HEAD> have molecular reaction of <TAIL> ?*
	part of	*Is <HEAD> part of <TAIL> ?*

Table 7: Question templates used in QA-Model (Section 4.1).

manually created the question templates for each dataset: CDR and HOIP.

C Prompts

Table 8 shows an example of the prompt used in the few-shot setting in process identification. Only examples section is discarded in the zero-shot setting. Table 9 also shows a prompt used in DocRE experiments on the HOIP dataset and the corresponding output by Llama3 (8B). We replaced the ontology name ("HOIP") and possible relation classes ("has-result, has-part, ...") in the prompt template with "MeSH" and "Chemical-Induce-Disease" respectively in CDR experiments. The demonstrations are also different between the datasets.

D ICL Few-Shot Setting in Process Identification

Table 10 exhibits the number of examples per cluster created for ICL in the few-shot setting in process identification.

E Hyperparameters

Table 11 shows the hyper-parameters used in the supervised models for process identification. Table 12 also list hyper-parameters used in our DocRE models.

F Another Example of Case Study

Table 13 shows another example used in our case study (in Section 7).

Instruction:	Generate the list of processes present in the Text.
Constraints:	Don't repeat the question. Justification and explanation are prohibited.
Examples:	**Text:** Within 19 days after symptom onset, 100% of patients tested positive for antiviral immunoglobulin-G (IgG). Seroconversion for IgG and IgM occurred simultaneously or sequentially. **Answer:** [immunoglobulin production, immunoglobulin mediated immune response]
Text:	ACE2 expression has been demonstrated in arterial and venous endothelium of several organs, and histopathological studies have found microscopic evidence of SARS-CoV-2 viral particles in endothelial cells of the kidneys and lungs.
Answer:	-

Table 8: Example of the few-shot setting prompt in process identification, following the described prompt template.

Prompt:

Based on the given text and entities associated with the text, please identify relations between the entities.
1. Named entities are listed next to the text.
2. Each entity is represented using HOIP Concept ID.
3. Possible relations: has-result, has-part, has-molecular-reaction, part-of
4. Output a bulleted list of triples. Each bullet line corresponds to each triple: "<BULLET> (<SUBJECT ENTITY>, <RELATION>, <OBJECT ENTITY>)",
where <SUBJECT ENTITY>, <RELATION>, and <OBJECT ENTITY>, correspond to the subject entity, the relation label, and the object entity, respectively.

Below are some examples.

Example 1
Text: We also provide biophysical and structural evidence that ...
Entities:
* http://purl.bioontology.org/ontology/HOIP/HOIP_0040511: Negative regulation of ACE2 activation
* http://purl.bioontology.org/ontology/HOIP/HOIP_0041139: host cell surface receptor binding in nasal epithelial cells
...
Answer:
1. (http://purl.obolibrary.org/obo/GO_0046789, has-result, http://purl.bioontology.org/ontology/HOIP/HOIP_0040511)
2. (http://purl.obolibrary.org/obo/GO_0046789, has-part, http://purl.bioontology.org/ontology/HOIP/HOIP_0041139)
...

Example 2
...
Example 3
...

Let's try the following test example.

Test Example
Text: Within 19 days after symptom onset, 100% of patients tested positive for antiviral immunoglobulin-G (IgG). Seroconversion for IgG and IgM occurred simultaneously or sequentially.
Entities:
* http://purl.obolibrary.org/obo/GO_0002377: immunoglobulin production
* http://purl.obolibrary.org/obo/GO_0016064: immunoglobulin mediated immune response

Please output the answer to the test example in bullet points, following the format specified above.

Generated response:

Here is the answer to the test example:

• (http://purl.obolibrary.org/obo/GO_0002377, has-result, http://purl.obolibrary.org/obo/GO_0016064)

Note that there is only one possible relation between the entities in this example, which is "has-result".

Table 9: Example of the whole prompt in DocRE and the corresponding Llama3 (8B) output.

Cluster	0	1	2	3	4	5	6	7	8	9
# Elements	3	9	4	4	1	5	3	1	4	1

Table 10: Number of elements for each cluster created for ICL few-shot setting. In cluster assignement, only clusters 0, 2, 3, 5, 6 appear in the HOIP test set

Hyper-parameter	
Max Sequence Length	512
Optimizer	AdamW
Learning Rate	1×10^{-5}
Weight Decay	1×10^{-6}
Epochs	8
Batch Size	8

Table 11: Hyper-parameters for the supervised models on Process Identification.

Hyper-parameter	QA-Model	MA-ATLOP	LLM-ICL
Pre-trained model	SciBERT (cased)	SciBERT (cased)	Llama3 (8B; instruction-fine-tuned)
Max Sequence Length	512	512	4096
Bilinear Group G	-	64	-
Negative Sampling Ratio ρ	-	0.5	-
Optimizer	AdamW	AdamW	-
Learning Rate (BERT encoders)	2×10^{-5}	2×10^{-5}	-
Learning Rate (FFNNs)	1×10^{-4}	1×10^{-4}	-
Epochs	30	30	-
Batch Size	4	2	-
Warmup Ratio	0.06	0.06	-
# Few-Shot Examples	-	-	3
Quantization Bits	-	-	4
dtype	-	-	BFloat16
Max. New Tokens	-	-	512

Table 12: Major hyper-parameters for the DocRE models.

ID242. Passage:
Moreover, isolated right ventricular dysfunction may occur as a result of elevated pulmonary vascular pressures secondary to ARDS, pulmonary thromboembolism, or potentially virus-mediated injury to vascular endothelial and smoothmuscle tissue.
True-Positive Entities:
* increasing blood pressure
* respiratory blood vessel smooth muscle damage
* thrombus formation
False-Negative Entities:
* artery narrowing
* endothelium damage
* endothelium malfunction
* vasoconstriction
True-Positive Triples:
* (endothelium damage, has result, endothelium malfunction)
* (thrombus formation, has result, artery narrowing)
* (vasoconstriction, has result, increasing blood pressure)
False-Negative Triples:
* (respiratory blood vessel smooth muscle damage, has result, vasoconstriction)

Table 13: A case study on our process identification and DocRE models.