CoastRED: A Corpus for Multidisciplinary Joint Entity and Relation Extraction in Coastal Scientific Literature

Anonymous ACL submission

Abstract

The increasing impact of climate change on coastal areas requires efficient collaboration among stakeholders to develop appropriate environmental protection policies. We introduce CoastRED, a corpus of 417 scientific abstracts focused on coastal regions for interdisciplinary Joint Entity and Relation Extraction (JERE) at the document level, crucial for Knowledge Base (KB) construction. Using the ARDI framework (Actors, Resources, Dynamics, and Interactions), our approach automatically extracts key entities and their roles in coastal systems. To overcome the limitations of sequential models for JERE, we present DUNJERE, an end-to-end model for joint Mention Detection, Coreference Resolution, Entity Typing, and Relation Extraction tasks. Inspired by DocuNet, DUNJERE is built on a U-Net-based architecture and treats Relation Extraction as a semantic segmentation task. Evaluated on the DocRED dataset, DUNJERE shows competitive performance, overcoming document-level challenges. Results on CoastRED demonstrate its potential for advancing coastal system research. Code and dataset will be publicly available on GitHub with the camera-ready version.

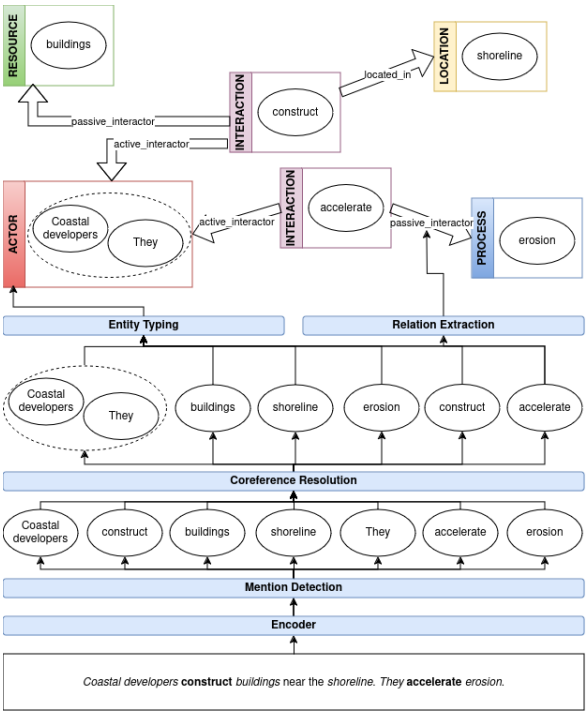

Figure 1: A high-level overview of DUNJERE's functionality on CoastRED. The model processes raw text inputs to identify entity mentions, clusters these mentions into distinct entities, classifies each entity into a specific type, and extracts the relationships between the identified entities. We treat Interactions as we would in Open Information Extraction (Niklaus et al., 2018).

1 Introduction

Coastal regions, impacted by both global environmental changes and human activities, represent complex systems characterized by the continuous interplay of physical, chemical, biological, and societal dynamics. Understanding these regions requires a holistic approach, acknowledging their status as inherently anthropized environments. Human interventions—such as coastal development, resource management, and urbanization—are integral to comprehending the various agents and mechanisms at play in these ecosystems.

This interdisciplinary nature presents a compelling case for joint entity and relation extraction, as coastal literature spans diverse fields, including environmental science, geography, ecology, and sociology. The ARDI framework (Etienne et al., 2011) offers a structured approach for identifying key elements in such systems by categorizing *Actors*, *Resources*, *Dynamics*, and *Interactions*. By emphasizing the system's functioning from an interdisciplinary perspective, it facilitates stakeholder collaboration in constructing a conceptual model of coastal systems, essential for developing effective environmental policies. However, the vast body of literature and the interdisciplinary breadth of the subject matter introduce significant challenges in accurately identifying and extracting relevant

entities and relationships.

The extraction of structured information from unstructured text is critical in Natural Language Processing (NLP), particularly for Knowledge Graph (KG) construction. This process typically involves several subtasks, including Mention Detection (MD), Coreference Resolution (CR), Entity Typing (ET), and Relation Extraction (RE). Traditional models often address these tasks sequentially, which can result in error propagation and diminished performance (Nasar et al., 2021). Recently, there has been a shift towards multi-task learning, where joint models handle theses tasks simultaneously, leveraging task interdependencies to improve performance and efficiency (Bekoulis et al., 2018; Wadden et al., 2019; Wang and Lu, 2020; Zhong and Chen, 2020; Yan et al., 2021; Zaratiana et al., 2024; Prieur et al., 2024).

Despite this progress, most current models are limited to sentence-level processing (Zheng et al., 2017; Bekoulis et al., 2018; Yan et al., 2021; Shang et al., 2022), restricting the scope of extracted information. Document-level extraction introduces additional challenges (Han and Wang, 2020): (1) a significantly larger number of entities, with the complexity of relationships increasing quadratically as entities grow, (2) the need for coreference resolution to account for multiple mentions of the same entity, (3) logical inference across multiple sentences, with around 40.7% of relations in DocRED (Yao et al., 2019), the reference dataset in the domain, requiring cross-sentence reasoning. Sentence-level approaches typically involve a separate forward pass for each sentence, leading to higher inference times and inefficiencies in practical applications (Bouziani et al., 2024). Only a few models have successfully tackled document-level JERE (DocJERE) (Zaporojets et al., 2021; Eberts and Ulges, 2021; Huguet Cabot and Navigli, 2021; Bouziani et al., 2024).

To address these challenges, we propose DUNJERE, a Document-level U-Net-based Network for Joint Entity and Relation Extraction. DUNJERE processes unstructured text by jointly extracting mentions, clustering them into entities, and identifying relations in an entity-centric manner. We evaluate DUNJERE on a benchmark dataset, DocRED-E2E (Bouziani et al., 2024), an adapted version of DocRED (Yao et al., 2019) introduced by Eberts and Ulges (2021) for DocJERE. This dataset provides extensive annotations for Named Entity Recognition (NER), CR and RE at document-level.

To summarize, our main contributions are :

- We introduce **CoastRED**, a multidisciplinary corpus for DocJERE on scientific abstracts related to coastal areas.
- We present **DUNJERE**, an end-to-end framework for entity-centric DocJERE inspired by DocuNet (Zhang et al., 2021).
- We demonstrate the effectiveness of DUNJERE through experiments on DocRED-E2E, showing competitive performance with state-of-the-art (SOTA) models, and test it on our CoastRED dataset.

This paper is organized as follows. We present related work in Section 2. Then, introduce our novel CoastRED dataset in Section 3, followed by Section 4, where we present DUNJERE architecture. The experiment setups are described in Section 5 before we highlight our general observation and findings in Section 6. A comprehensive error analysis is shown in Section 7. Finally, in Section 8, we present our conclusions and future work, and discuss the limitations of the proposed methods in Section 9.

2 Related Work

2.1 Information Extraction in Environmental Corpora

Efforts to extract structured information from environmental corpora have led to the development of several KBs and tools. TermFrame[1] serves as a KG constructed by extracting terms and triplets from English and Slovene corpora focusing on karst environments (Pollak et al., 2019; Vintar and Martinc, 2022). Similarly, EcoLexicon (Faber et al., 2013, 2016) is a multilingual KB specialized in environmental concepts, covering six languages (English, French, German, Modern Greek, Russian, and Spanish). For NER tasks in domain-specific contexts, TaxoNERD (Le Guillarme and Thuiller, 2022) targets taxonomic mention recognition in ecological texts, while AGRONER (Veena et al., 2023) applies unsupervised NER techniques to agricultural texts.

2.2 Sentence-level Joint Entity and Relation Extraction

Recent advancements in end-to-end models for JERE seek to mitigate the cascading errors inherent

[1] https://termframe.ff.uni-lj.si/

in sequential processing. These models enhance task interaction and performance by optimizing entity and relation extraction together. Fu et al. (2019) introduced GraphRel, which leverages Graph Convolutional Networks (GCNs) to jointly learn named entities and relations. Luan et al. (2019) developed the DyGIE model, which constructs dynamic span graphs to perform NER, RE, and CR tasks simultaneously . Other models, such as Sun et al. (2021), progressively learn tasks by utilizing lower-layer predictions in a multitask framework , and Ye et al. (2021) use levitated markers to encode span pairs strategically. Techniques such as table-filling (Wang and Lu, 2020) and 2D convolutional networks on tables (Ma et al., 2022) further enhance local dependency modeling in these tasks. Other approaches like Sui et al. (2023) focus on direct set prediction, avoiding the complexities of multiple triple prediction, while models like UniRE (Wang et al., 2021) and UTC (Yan et al., 2023) exploit the unified label space or token-pair classification for joint extraction. However, Enrico (Zaratiana et al., 2024) take an attention-based approach, dynamically identifying relevant information for entity and relation extraction. Despite their effectiveness, sentence-level methods miss cross-sentence triples and are computationally expensive, as inference time scales linearly with the number of sentences (Bouziani et al., 2024).

2.3 Document-level Joint Entity and Relation Extraction

Several models have been proposed to integrate NER, CR, and Document-level Relation Extraction (DocRE) tasks within a unified framework. JEREX (Eberts and Ulges, 2021) employs multi-instance learning, while DWIE (Zaporojets et al., 2021) utilizes a Graph Neural Network (GNN) to propagate local contextual mention span information. REBEL (Huguet Cabot and Navigli, 2021) uses an autoregressive sequence-to-sequence model, leveraging additional data such as Wikidata to enhance relation extraction performance. Similarly, Verlinden et al. (2021) incorporate entity representation from Wikipedia and Wikidata to improve information extraction performance. Additionally, REXEL (Bouziani et al., 2024) performs Entity Disambiguation (ED), linking extracted entities to a KB.

2.4 Document-level Relation Extraction

Methods for DocRE typically fall into two categories: Transformer-based and Graph-based approaches. Transformer-based methods include models like ATLOP (Zhou et al., 2020), which employs adaptive thresholding to optimize entity pair classification and localized context pooling for contextual information extraction. DREEAM (Ma et al., 2023) enhances this by generating context embeddings specific to entity pairs and incorporating attention weights as evidence scores, improving evidence extraction without increasing computational costs. Alternatively, DocuNet (Zhang et al., 2021) uses a U-Net-based architecture (Ronneberger et al., 2015), applying semantic segmentation techniques to model entity interdependencies. DRE-MIR (Zhang and Cheng, 2022) incorporates a random masking strategy during training to reconstruct the entity-pair matrix during inference. AutoRE (Xue et al., 2024) leverages Large Language Models (LLMs) combined with QLoRA (Dettmers et al., 2024) to perform the DocRE task. Lastly, DocRE-CLiP (Jain et al., 2024) reformulates the task as a KG-based link prediction problem to incorporate external entity context.

Graph-based approaches, on the other hand, construct various graph structures to capture entity relationships. GAIN (Zeng et al., 2020) employs a dual-graph architecture, consisting of mention-level and entity-level graphs to capture interactions. HAG (Li et al., 2022) combines a coreference graph with a mention graph, using a self-supervised affinity computation mechanism. EGCN (Sun et al., 2022) integrates node importance metrics from PageRank (Page et al., 1997) to improve the model's ability to capture the topological structure of entity relationships.

3 CoastRED Dataset

3.1 Annotation

In this section, we detail the annotation process of CoastRED.

3.1.1 Overall Process

We collected a corpus of 64,000 papers from Scopus[2] (1980–2023) containing the terms "*coastal areas*" or "*littoral*" in their titles or abstracts. From this, we randomly selected 600 abstracts, comprising 60% research articles and 40% surveys, for manual annotation. The annotation was carried out by two Master's students in earth sciences, who were compensated and provided with annotation

[2] https://www.elsevier.com/fr-fr/solutions/scopus

guidelines, developed by a Ph.D. student in computer science and coastal research, along with a domain expert and ontology specialist. Only sentences that contributed to understanding coastal zone dynamics were annotated, while methodological descriptions were excluded. Two versions of the guidelines were created: a detailed one with figures and examples, and a condensed version with essential instructions.

Fourteen relation labels, mainly derived from the Relation Ontology[3], were also defined and are detailed in Appendix A. Annotation was conducted by the two students, the domain expert, and the Ph.D. student using the INCEpTION[4] tool (Klie et al., 2018). The campaign aimed for dual-annotator coverage for 60% of the abstracts. In total, 215 abstracts were annotated, with a Krippendorff's alpha (Krippendorff, 2011) of 43%, reflecting moderate agreement but falling short of the 66% threshold for direct dataset use. The annotations were then curated to ensure consistency.

In parallel, we utilized three domain-specific knowledge bases (AGROVOC[5], GEMET[6], AFO[7], and TAXREF-LD[8]) to pre-annotate 195 abstracts. We then refined these annotations using the same guidelines, ultimately harmonizing the manually and KB-annotated datasets for coastal area research. The two subsets were subsequently combined to create a train-validation-test split with 266 abstracts in the train set, 67 in the validation set and 84 in the test set.

3.1.2 Annotation of Interactions

For annotating interactions, we considered several approaches: [1] *Treat them as relations with an exhaustive list*: annotators would select from a predefined list of interactions. However, this method is time-consuming, limits flexibility for identifying new interactions, and would require retraining the model for future updates; [2] *Treat them as relations with a flexible list*: A less exhaustive list would allow annotators to create new labels when necessary, with synonyms merged during curation. Although more efficient, this approach still poses challenges, such as synchronizing the relation list across annotators and requiring re-annotation and retraining for identifying new interactions identified in scientific articles; [3] *Annotate the spans with an "Interaction" label*: This method involves creating two relations— "*passive_interactor*" and "*active_interactor*"—similar to Open Information Extraction. It enables capturing a broad range of interactions without predefined labels and allows the model to detect new interactions during inference. While it may introduce ambiguity in identifying text spans, this approach is more adaptable and could be refined by mapping interaction spans to a relation ontology. Additionally, this more permissive approach enables the linking of interactions with other relations, such as providing contextual information (e.g., location - where the interaction occurs).

Given these considerations, we opted for the third approach. We also introduce a third relation- "*intermediate_interactor*"-to handle n-ary interactions, for instance when another entity is used by the active interactor to perform the interaction. We follow the same setting for the "*Comparison*" label, with two relations, namely "*compared*" and "*compared_to*".

3.2 Data Analysis

In this section, we conduct a comprehensive analysis of various aspects of CoastRED to enhance our understanding of both the dataset and the task at hand.

3.2.1 Data Size

Table 1 provides a comparative overview of CoastRED and two representative DocJERE datasets, namely DocRED-E2E (Bouziani et al., 2024) and DWIE (Zaporojets et al., 2021). While CoastRED is smaller in size compared to the other datasets, it features a proportionally higher number of entity mentions and relations. On average, each abstract in CoastRED comprises 9.5 sentences. Even when mentions labeled as "*Interactions*" and "*Comparison*" are excluded, CoastRED still retains a substantial 25,127 mentions, with 38.7% of the mentions being nested within at least one other, and 11,132 unique terms (lemmatized) on a total of 23,103 terms. The dataset's type-to-token ratio stands at 0.11. Additionally, since only sentences contributing to the understanding of coastal area dynamics were annotated, 1,458 sentences in the dataset do not contain any annotated mentions above all 3,971 sentences.

[3] https://www.ebi.ac.uk/ols/ontologies/ro
[4] https://inception-project.github.io/
[5] https://agrovoc.fao.org/browse/agrovoc/en/
[6] https://www.eionet.europa.eu/gemet/en/about/
[7] https://finto.fi/afo/en/
[8] https://github.com/frmichel/taxref-ld

Datasets	# Documents	# Tokens	Entities			Relations	
			# Mentions	# Entity clusters	# Entity types	# Relations	# Relation types
DocRED-E2E	4,008	794,729	104,367	78,107	6	49,951	96
DWIE	802	501,095	43,373	23,130	311	21,749	65
CoastRED	417	107,106	25,127	18,317	8	23,070	15

Table 1: Numerical comparison of CoastRED and other popular datasets in DocJERE

3.2.2 Entity Types

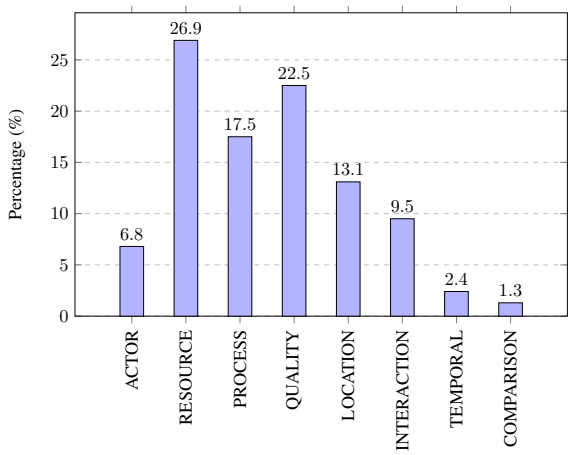

Figure 2: Proportions of entity types in the dataset.

We adapted the ARDI framework (Etienne et al., 2011) to extract system-related information, defining the following labels: "*Actor*" (stakeholders initiating processes or consuming resources), "*Resource*" (goods, elements, or facilities used by stakeholders), "*Process*" (environmental, societal, or economic processes, as defined by the Environment Ontology (Buttigieg et al., 2013)), "*Quality*" (e.g., height or concentration, as defined by PATO (Gkoutos et al., 2017)), "*Location*" (geographical areas and landmasses, bodies of water, and geological formations), "*Temporal*" (date, or temporal item), "*Interaction*", and "*Comparison*". As illustrated in Figure 2, there is a predominance of "*Resource*" and "*Quality*" entities, while "*Actor*" is comparatively underrepresented.

3.2.3 Relation Types

The number of relation labels is deliberately kept limited, as the more specific interactions are captured through the "*Interactions*" framework. Consequently, we employ a set of fundamental relations, largely inspired by the Relation Ontology[9] (based on the Basic Formal Ontology[10]). These include, in addition to the "*active*", "*passive*" and "*intermediate_interactor*" as well as "*compared*" and "*compared_to*": "*influences*", "*cause*", "*part_of*", "*located_in*", "*has_attribute*", "*has_value*", "*bearer_of*", "*temporal_specification*", "*frag*".

[9] https://www.ebi.ac.uk/ols/ontologies/ro
[10] https://www.ebi.ac.uk/ols/ontologies/bfo

4 DUNJERE Architecture

4.1 Encoding

We feed each document $d = [x_t]_{t=1}^{l}$ into a pre-trained language model to obtain the contextual embeddings:

$$H = [h_1, h_2, ..., h_l] = \texttt{BERT}([x_1, x_2, ..., x_l])$$

Following ATLOP (Zhou et al., 2020), for documents exceeding 512 tokens in length, we employ a dynamic windowing technique which averages the embeddings of overlapping tokens of different windows to encode the entire document.

4.2 Mention Detection

We first compute mention representation by enumerating all possible spans from the input sentence (up to a maximum span length in practice):

$$S_{ab} = w_{ent}^{T}(h_a \oplus h_b)$$

S_{ab} the span representation for a span starting at word a and ending at word b, $\oplus$ denoting the concatenation operation, and h_a the embedding of the first token and h_b the one of the last token of the entity mention. Eventually, we compute the span classification score as part of a binary classifier.

4.3 Coreference Resolution

We reformulate the CR task as a RE problem, classifying each mention pair as either coreferent or non-coreferent. To implement this, we compute the cosine similarity between mention embeddings and apply a linear transformation similar to standardization. Rather than employing traditional mean-based standardization—which can lead to inconsistencies due to batch-wise variations in cosine magnitudes—we use a linear transformation with a

learnable threshold. This approach provides greater adaptability and stability in the model. The transformed values are then fed into a linear layer for the final classification.

4.4 Entity Typing

We add two coreferent mentions to the same entity cluster, enabling a chained process to include all positions of the entity's mentions within each cluster. The entity embedding e_i is then obtained by applying $logsumexp$ pooling to the embeddings of its mentions (Jia et al., 2019). We pass the entity embeddings via a linear block for classification.

4.5 Relation Extraction

Inspired by DocuNet (Zhang et al., 2021), we apply a U-shaped semantic segmentation module to the RE task. Entity embeddings are processed through this module. We leverage a context-based method using entity-aware attention with affine transformation to obtain the feature vector as follows:

$$F(e_i, e_j) = wHa^{(s,o)}$$

$$a^{(i,j)} = \texttt{softmax}(\sum_{k=1}^{K} A_k^i A_k^j)$$

where $a^{(i,j)}$ is the attention weight for entity-aware attention and A_k^i refers to the tokens' importance to the k-th entity. The matrix is processed through a U-shaped segmentation architecture comprising two down-sampling and two up-sampling blocks, linked via skip connections. Each down-sampling block consists of max-pooling followed by convolutional modules, with the number of channels doubling at each stage. This design enhances interaction between entity pairs by increasing the receptive field, thereby improving the incorporation of global information for more robust representation learning. In the up-sampling blocks, deconvolution and convolution modules are employed, progressively halving the number of channels at each stage to effectively distribute the aggregated information across pixels. To improve computational efficiency, depthwise separable convolutions (Sandler et al., 2018) are used in place of conventional convolutions.

Given the entity pair embedding e_i and e_j and the entity-level relation matrix Y, we map them to hidden representations z with a feedforward neural network. The relation probability is then computed using a bilinear function. Formally, this is described as follows:

$$z_i = tanh(W_i e_i + Y_{i,j})$$

$$z_j = tanh(W_j e_j + Y_{i,o})$$

$$P(r|e_i, e_j) = \sigma(z_i W_r z_j + b_r)$$

Here, $Y_{i,j}$ denotes the entity-pair representation of (i, j) in matrix Y. $W_r \in R^{d\times d}$, $b_r \in R$, $Ws \in R^{d\times d}$, and $W_j \in R^{d\times d}$ are learnable parameters.

To address the class imbalance common in RE task—where most entity pairs do not share a relation (Wang et al., 2019) and thus are attributed the "*Na*" relation label—we use the loss function as employed in DocuNet for the CR and RE modules (Su, 2020):

$$\mathcal{L} = log(1 + \sum_{k\in\omega_{\text{neg}}} e^{i_k}) + log(1 + \sum_{k\in\omega_{\text{pos}}} e^{j_k})$$

4.6 Optimization and Inference

We optimize the joint training by using a weighted sum of the module-specific losses, with fixed weights that serve as tunable hyperparameters, as follows:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{MD} + \beta\mathcal{L}_{CR} + \gamma\mathcal{L}_{ET} + \varphi\mathcal{L}_{RE}$$

During training, ground truth and gold inputs are provided for all subtasks. When training on a single subtask, the weights for all other task losses are set to zero. For end-to-end inference, we utilize predicted mention spans and predicted entities as input.

5 Experiments

5.1 Datasets

We evaluate DUNJERE using data from DocRED (Yao et al., 2019), which consists of gold-annotated documents sampled from Wikipedia. The dataset's NER component features traditional entity labels such as "*Person*", "*Organization*", and "*Location*", while its RE component includes 96 relation types spanning various domains such as science, art, time, and personal life. Notably, the labels for the original test set are hidden, and evaluation for DocRE requires the submission of results through Codalab[11]. This led Eberts and Ulges (2021) to establish a new data split by merging the training and development sets to assess end-to-end systems. They randomly

[11] https://competitions.codalab.org/competitions/20717

sampled a training set of 3,008 documents, a development set of 300 documents, and a test set of 700 documents, removing 45 documents due to incorrect annotations involving entities of different types. This set was augmented with Wikidata links on entities for the Entity Disambiguation task by Bouziani et al. (2024) to make the DocRED-E2E dataset.

5.2 Subtasks

We report the performance for each subtask independently, setting all coefficients (α, β, γ, φ) to zero except for the one corresponding to the subtask in question. Additionally, we evaluate the model in a joint learning context, both under conditions where the ground truth is provided to each module and where each module is fed with the outputs of the preceding module. This approach enables us to evaluate the model's performance in isolated subtask settings, as well as within the integrated pipeline, and to determine the potential synergistic benefits between tasks.

5.3 Evaluation Metrics

We evaluate the model's performance using F1-score (F1) across each subtask. In MD task, a predicted mention is considered correct if its boundaries precisely match the ground truth. For CR task, a cluster is considered correct if it contains exactly the same mentions as those in the ground truth. In end-to-end RE task, a triple is classified as correct if the predicted relation type and the entity clusters corresponding to the head and tail entities are both accurate (Eberts and Ulges, 2021; Zaporojets et al., 2021; Bouziani et al., 2024).

6 Results

We conducted a comparative analysis of our results against two other approaches on DocRED-E2E, JEREX (Eberts and Ulges, 2021) and REXEL (Bouziani et al., 2024), and assessed the performance of DUNJERE using three backbones, RoBERTa (Liu et al., 2019), SciBERT (Beltagy et al., 2019) and DeBERTa (He et al., 2021), on the CoastRED dataset as summarized on Table 2.

6.1 DocRED-E2E

To evaluate the performance of DUNJERE's components, each subtask was trained and evaluated separately using the DocRED-E2E dataset. In isolated task training, DUNJERE surpasses SOTA in MD and ET, while delivering competitive performance in CR and Relation Classification (RC). When trained jointly across all modules, DUNJERE outperforms REXEL in end-to-end RE. However, the performance of MD and CR appears to decline, potentially due to differences in loss magnitudes between these tasks and ET, and the distributed focus of the training effort across all subtasks.

6.2 CoastRED

In isolated task training, CR and ET demonstrate strong performance, even if we denote a lower performance with DeBERTa as a backbone, probably due to overfitting . However, MD exhibits significantly lower performance compared to results on DocRED, which is expected, as MD on CoastRED involves term extraction rather than named entity recognition, making the task inherently more challenging. Additionally, the low performance in RC underscores the complexity of this task within the dataset.

7 Error analysis

7.1 Impact of sentences without mentions

As can be seen, the empty sentences harm the model's performance on MD (almost a 10 % augmentation in F1 when we remove the sentences without any annotated mentions). However, in the end-to-end setting, this drop isn't followed then in CR, since this module obtains comparable results in both settings, although it obtains less entity mentions as inputs. We can formulate the hypothesis that with the empty sentences still in the dataset, the module receives the easiest mentions to clusterize. More importantly, except for DeBERTa, keeping the empty sentences seems to ameliorate the results of the RC module in the joint setting, while ET doesn't seem to be affected.

7.2 Impact of mention type on mention detection

To evaluate the influence of mention type on detection accuracy, we present the results by mention type in Figure 3. The mentions labelled as "*Actor*", "*Resource*", "*Process*", and "*Location*" are identified with relatively high accuracy (between 65% and 80%), while "*Quality*" mentions show slightly lower performance (around 60%) for RoBERTa, and even lower for SciBERT and DeBERTa (a little more than 50%). This is likely due to the greater

Models	Subtasks				Subtasks simultaneously				End-to-end			
	MD	CR	ET	RC	MD	CR	ET	RC	MD	CR	ET	RC
DocRED-E2E												
JEREX	92.66	90.46	95.29	59.76	**92.99**	82.79	80.10	/	/	/	/	**40.38**
REXEL	90.56	**90.93**	96.01	**60.10**	90.68	**89.02**	95.78	57.38	/	/	/	39.06
DUNJERE (RoBERTa$_{base}$)	**92.81**	89.74	**97.24**	57.52	90.84	88.68	**97.35**	57.52	/	85.06	78.95	39.14
CoastRED												
DUNJERE (RoBERTa$_{base}$)	64.50	**85.66**	90.73	35.10	63.58	**82.96**	89.13	32.05	/	**65.35**	**50.86**	**12.39**
DUNJERE (SciBERT)	65.43	84.04	91.18	38.27	62.24	69.41	**89.42**	31.21	/	49.71	39.19	8.31
DUNJERE (DeBERTa-v3$_{base}$)	**66.03**	77.24	**91.42**	**40.70**	**63.93**	71.13	88.84	**37.30**	/	53.32	39.73	8.60
CoastRED without non-mention sentences												
DUNJERE (RoBERTa$_{base}$)	73.57	**85.17**	90.81	38.47	71.47	**84.05**	89.65	28.22	/	**66.97**	**52.59**	**12.02**
DUNJERE (SciBERT)	74.29	84.23	91.11	40.04	71.99	68.29	89.67	27.26	/	52.12	43.09	9.41
DUNJERE (DeBERTa-v3$_{base}$)	**75.01**	78.29	**91.38**	**41.42**	**73.77**	72.62	**90.14**	**39.32**	/	59.13	45.65	10.57

Table 2: Evaluation in F1 of DUNJERE on different datasets. The best score on a subtask is written in **bold** and the second best is underlined. "Subtasks" refers to training the model for individual tasks, "Subtasks simultaneously" involves joint training across all tasks with each module receiving ground truth inputs, and "End-to-End" indicates the same joint training, but during inference, the model receives the outputs of the preceding module instead of ground truth inputs.

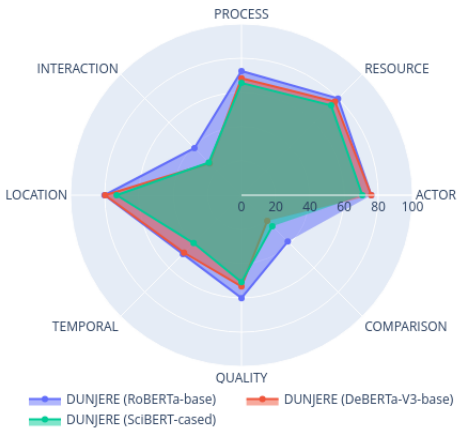

Figure 3: DUNJERE's MD accuracy on CoastRED, categorized by mention type, for a joint training.

complexity and variability associated with "*Quality*" mentions. However, "*Interaction*" and "*Comparison*" types exhibit notably lower detection accuracy for all models, highlighting the significant challenges in identifying these categories due to their particularity of being closer to a RE task than MD.

7.3 Relation Extraction

DUNJERE's performance on DocRE for CoastRED is notably poor. During preprocessing, we introduce an additional "*Na*" label for entity pairs that do not share a relation, leading to a classification task with $n + 1$ labels, where n corresponds to the number of annotated labels. However, the vast majority of entity pairs, specifically 97.38%, are assigned the "*Na*" label, resulting in highly imbalanced classes. This makes it exceedingly challenging for the model to accurately identify true relations, as each relation accounts for less than 1% of the total cases. Although DocRED-E2E faces a similar imbalance (with 96.77% of entity pairs labeled as "*Na*"), the relations in CoastRED are more complex, often requiring domain-specific knowledge about the system, which further complicates the task.

8 Conclusion

In this work, we introduced CoastRED, a novel corpus for DocJERE within the context of coastal systems research. Rather than utilizing traditional entity-type labels, we employed the ARDI framework (Etienne et al., 2011) to structure the extraction process, emphasizing the functional roles of entities within coastal systems. To evaluate this dataset, we presented DUNJERE, a Document-level U-Net-based Network for Joint Entity and Relation Extraction. Our evaluations on DocRED-E2E demonstrate competitive performance across tasks. However, the results for RC on CoastRED indicate significant room for future improvement. These findings highlight the potential of DUNJERE in advancing DocJERE methodologies while pointing to areas for further refinement.

9 Limitations

CoastRED Dataset. CoastRED is relatively small, despite the richness of information it contains, which may lead to early overfitting. Expanding the dataset is not straightforward, as creating new annotated data is time-consuming due to the complexity of the annotations, including nested mentions and intricate interactions and relations. Furthermore, the model struggles to accurately identify "*Interactions*" and relations, which are crucial for understanding the coastal systems described in the documents.

DUNJERE Architecture. Unlike other models such as REXEL, DUNJERE currently lacks an Entity Disambiguation module. It also faces constraints in context length, limited by the maximum number of tokens that can be processed by its encoder. However, this issue is largely mitigated in the case of both DocRED-E2E and CoastRED, as sufficient coreferences exists to reduce the distance between entities. Moreover, there is evidence of gradient dilution across tasks, especially when trained on CoastRED. This might suggest that training strategies that prioritize sequential training of modules, followed by joint training, may offer a more effective and efficient approach than training all tasks simultaneously from the start.

References

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann, and Andrea Pierleoni. 2024. Rexel: An end-to-end model for document-level relation extraction and entity linking. *arXiv preprint arXiv:2404.12788*.

Pier Luigi Buttigieg, Norman Morrison, Barry Smith, Christopher J Mungall, Suzanna E Lewis, and Envo Consortium. 2013. The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics*, 4:1–9.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. *arXiv preprint arXiv:2102.05980*.

Michel Etienne, Derick R Du Toit, and Sharon Pollard. 2011. Ardi: a co-construction method for participatory modeling in natural resources management. *Ecology and society*, 16(1).

Pamela Faber, Pilar León-Araúz, and Arianne Reimerink. 2013. Representing environmental knowledge in ecolexicon. In *Languages for specific purposes in the digital era*, pages 267–301. Springer.

Pamela Faber, Pilar León-Araúz, and Arianne Reimerink. 2016. Ecolexicon: new features and challenges. *GLOBALEX*, pages 73–80.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. Graphrel: Modeling text as relational graphs for joint entity and relation extraction. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1409–1418.

Georgios V Gkoutos, Paul N Schofield, and Robert Hoehndorf. 2017. The anatomy of phenotype ontologies: principles, properties and applications. *Briefings in Bioinformatics*, 19(5):1008–1021.

Xiaoyu Han and Lei Wang. 2020. A novel document-level relation extraction method based on bert and entity information. *IEEE Access*, 8:96912–96919.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. 2024. Revisiting document-level relation extraction with context-guided link prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18327–18335.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. *arXiv preprint arXiv:1904.02347*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Nicolas Le Guillarme and Wilfried Thuiller. 2022. Taxonerd: deep neural models for the recognition of taxonomic entities in the ecological and evolutionary literature. *Methods in Ecology and Evolution*, 13(3):625–641.

Rongzhen Li, Jiang Zhong, Zhongxuan Xue, Qizhu Dai, and Xue Li. 2022. Heterogenous affinity graph inference network for document-level relation extraction. *Knowledge-Based Systems*, 250:109146.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. *arXiv preprint arXiv:1904.03296*.

Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. 2022. Joint entity and relation extraction based on table labeling using convolutional neural networks. In *Proceedings of the Sixth Workshop on Structured Prediction for NLP*, pages 11–21, Dublin, Ireland. Association for Computational Linguistics.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Comput. Surv.*, 54(1).

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. A survey on open information extraction. *arXiv preprint arXiv:1806.05599*.

Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1997. Pagerank: Bringing order to the web. Technical report, Stanford Digital Libraries Working Paper.

Senja Pollak, Andraž Repar, Matej Martinc, and Vid Podpečan. 2019. Karst exploration: extracting terms and definitions from karst domain corpus. *Proceedings of eLex*, 2019:934–956.

Maxime Prieur, Cédric Du Mouza, Guillaume Gadek, and Bruno Grilheres. 2024. Shadowfax: Harnessing textual knowledge base population. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2796–2800.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11285–11293.

Jianlin Su. 2020. Extend "softmax+ cross entropy" to multi-label classification problem.

Dianbo Sui, Xiangrong Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Joint entity and relation extraction with set prediction networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2021. Progressive multi-task learning with controlled information flow for joint entity and relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13851–13859.

Qi Sun, Kun Zhang, Kun Huang, Xun Li, Ting Zhang, and Tiancheng Xu. 2022. Enhanced graph convolutional network based on node importance for document-level relation extraction. *Neural Computing and Applications*, 34(18):15429–15439.

G Veena, Vani Kanjirangat, and Deepa Gupta. 2023. Agroner: An unsupervised agriculture named entity recognition using weighted distributional semantic model. *Expert Systems with Applications*, 229:120440.

Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2021. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1952–1957, Online. Association for Computational Linguistics.

Špela Vintar and Matej Martinc. 2022. Framing karstology: From definitions to knowledge structures and automatic frame population. *Terminology*, 28(1):129–156.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online. Association for Computational Linguistics.

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. UniRE: A unified label space for entity relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 220–231, Online. Association for Computational Linguistics.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. AutoRE: Document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 211–220, Bangkok, Thailand. Association for Computational Linguistics.

Hang Yan, Yu Sun, Xiaonan Li, Yunhua Zhou, Xuanjing Huang, and Xipeng Qiu. 2023. UTC-IE: A unified token-pair classification architecture for information extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4096–4122, Toronto, Canada. Association for Computational Linguistics.

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. *arXiv preprint arXiv:2108.12202*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2021. Packed levitated marker for entity and relation extraction. *arXiv preprint arXiv:2109.06067*.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Urchade Zaratiana, Nadi Tomeh, Yann Dauxais, Pierre Holat, and Thierry Charnois. 2024. Enrico: Enriched representation and globally constrained inference for entity and relation extraction. *arXiv preprint arXiv:2404.12493*.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1630–1640, Online. Association for Computational Linguistics.

Liang Zhang and Yidong Cheng. 2022. A masked image reconstruction network for document-level relation extraction. *Preprint*, arXiv:2204.09851.

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.

Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66.

Zexuan Zhong and Danqi Chen. 2020. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2020. Document-level relation extraction with adaptive thresholding and localized context pooling. *Preprint*, arXiv:2010.11304.

A Relations Proportions

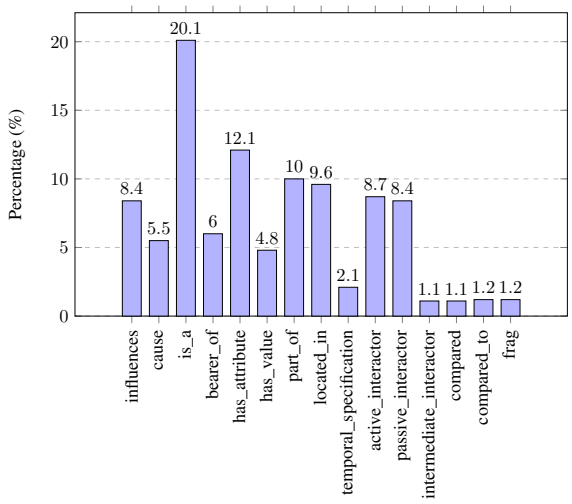

Figure 4: Proportions of annotated relation labels in CoastRED.

Figure 4 illustrates the distribution of relation labels in CoastRED. The most prevalent label, *"is_a"*, constitutes 20% of the annotated relations, followed by *"has_attribute"* (12.1%), *"part_of"* (10%), and *"located_in"* (9.6%). Labels such as *"influences"*, *"active_interactor"*, and *"passive_interactor"* are also well represented. In contrast, important relations like *"cause"* and *"bearer_of"* are underrepresented, despite their significance for describing coastal systems. Less frequent labels, including *"intermediate_interactor"*, *"compared"*, *"compared_to"*, and *"frag"*, account for around 1% of the annotations. This imbalance in label distribution may present challenges for models trained on CoastRED, as they are likely to struggle with accurately predicting infrequent relations.

B Training and implementation details

DUNJERE utilizes model implementations from the HuggingFace library[12] as its backbone architecture. The implementation of the model was carried out with the assistance of GitHub Copilot[13]. Optimization is carried out using AdamW (Loshchilov, 2017), with the learning rate reduced by a factor of 10 when the F1 score for the objective fails to improve after a specified number of patience epochs. The model's hyperparameters are detailed in Table 3. With RoBERTa$_{base}$ as a backbone, DUNJERE comprises 134M parameters. For joint training, the model was trained for up to 50 epochs on DocRED-E2E and 150 epochs on CoastRED, while for single-task training, it was capped at 50 epochs on DocRED-E2E and 80 epochs on CoastRED. All training procedures were conducted on a machine equipped with two NVIDIA H100 GPUs.

	Dataset	
Hyperparameter	DocRED-E2E	CoastRED
α	1	10
β	1	5
γ	0.1	0.05
φ	1	1
MD Learning rate	2.10^{-5}	2.10^{-5}
CR Learning rate	2.10^{-3}	2.10^{-3}
ET Learning rate	2.10^{-5}	2.10^{-5}
RE Learning rate	2.10^{-5}	2.10^{-3}
Batch size	4	4
Max span width	15	15
Patience	5	10

Table 3: DUNJERE model hyperparameters for the training processes.

[12] https://huggingface.co/models
[13] https://github.com/features/copilot