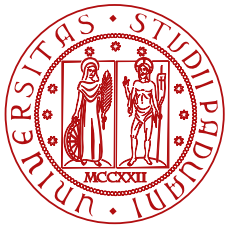

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in Computer Engineering

Entity and Relation Extraction for Marine Biology Research Papers

Master Candidate

Odai Mohammad

Supervisor

Giorgio Maria Di Nunzio

Date : 10 July

Academic Year

2023/2024

I dedicate this to my cat.

Abstract

Entity and Relation Extraction for Marine Biology Research Papers

ITALIANO: Entity and Relation Extraction for Marine Biology Research Papers

Contents

List of Figures

List of Tables

List of Algorithms

List of Code Snippets

List of Acronyms

IE Information Extraction

NER Named Entity Recognition

RE Relation Extraction

ACE The Automatic Content Extraction

FFN Feedforward Netwrok

KB Knowledge Base

LDC Linguistic Data Consortium

EDT Entity Detection and Tracking

LNK Entity Linking

RDC Relation Detection and Characterization

EDC Event Detection and Characterization

GPEs Geo-Political Entities

AI Artificial Intelligence

NYT dataset New York Times Relation Extraction Dataset

TACRED Text Analysis Conference Relation Extraction Dataset

TAC Text Analysis Conference

KBP Knowledge Base Population

NIST National Institute of Standards and Technology

LSTM Long Short-Term Memory

GCN Graph Convolutional Network

MRC Machine Reading Comprehension

DYGIE Dynamic Graph Interaction Extraction

BERT Bidirectional Encoder Representations from Transformers

AT Adversarial Training

PURE the Princeton University Relation Extraction system

1

Introduction

2

Related Work

With the rise of the Internet, there has been a notable surge in digital text creation across various platforms such as social media, emails, blogs, news articles, publications, and online forums. This vast corpus of unstructured or semi-structured text harbors a wealth of information. Information Extraction (IE) is a pivotal tool in discerning and organizing meaningful insights from these textual sources, transforming them into structured data.

One way to represent information in text is in the form of entities and relations representing links between entities. Therefore, Named Entity Recognition (NER) and Relation Extraction (RE) emerge as particularly valuable techniques and key components of IE. They enable extracting pertinent entities and relationships within the text, facilitating the conversion of raw data into structured repositories of valuable information.

The NER task identifies entities from the text, and the RE task can identify relationships between those entities. Furthermore, end-to-end relation extraction aims to identify named entities and extract relations between them in one go. They are effectively modeling these two subtasks jointly[1], either by casting them in one structured prediction framework or performing multi-task learning through shared representations.

Many NLP applications can benefit from relational information derived from natural language[2], including Structured Search, Knowledge Base (KB) population, Information Retrieval, Question-Answering, Language Understanding, Ontology Learning, etc. Therefore these tasks have been studied extensively and many datasets have been created and many models have been proposed to

tackle them.

2.1 NER and RE challenges

The NER and RE tasks face many challenges that need to be overcome. These challenges include and are not limited to:

- Domain-Specific Terminology and Context[3]: Adapting models to effectively handle domain-specific terminology, especially in specialized fields requires significant tuning and domain knowledge.
- Variability and Ambiguity in Text[4]: The inherent variability and ambiguity in natural language make it challenging to accurately identify and classify entities and relations, particularly in cases of sparse or implicit information.
- Data Scarcity and Annotation Quality: High-quality, annotated datasets are crucial for training effective models. However, the scarcity of such datasets in specific domains and the variability in annotation quality can hinder model performance and generalization.
- Cross-Domain and Cross-Linguistic Applicability: Developing models that perform well not only across different domains but also across languages is a significant challenge[5], requiring robust and adaptable methodologies.
- Integration of Knowledge Bases and External Information: Effectively integrating external knowledge bases and contextual information to improve the accuracy of NER and RE tasks remains a complex challenge[6].

2.2 Datasets

The exploration and understanding of complex textual data have significantly advanced with the development of NER and RE technologies. Central to these advancements are the diverse datasets that have been meticulously curated to train and evaluate these information extraction systems.

2.2.1 The Automatic Content Extraction dataset

The Automatic Content Extraction (ACE) Program was launched to boost the creation of technologies for processing language data automatically[7]. This

included tasks like classifying, filtering, and choosing data based on its content and the meanings conveyed. The main aim of the ACE Program was to improve technologies that could automatically recognize and describe these meanings, helping to enhance how machines understand natural language.

Central to the ACE Program were its research objectives: the detection and characterization of Entities, Relations, and Events. These objectives were meticulously addressed through the development of annotation guidelines, corpora, and other linguistic resources by the Linguistic Data Consortium (LDC), some in cooperation with the TIDES Program for supporting TIDES Extraction evaluations. The datasets produced under ACE, encompassing broadcast transcripts, newswire, and newspaper data in English, Chinese, and Arabic, became pivotal resources for training and testing in common research task evaluations.

The primary ACE annotation tasks were Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), Event Detection and Characterization (EDC), and Entity Linking (LNK). EDT laid the groundwork by identifying entities within a document across mentions—named, nominal, or pronominal. Entities were classified into seven types—Person, Organization, Location, Facility, Weapon, Vehicle, and Geo-Political Entities (GPEs), with further distinctions into subtypes. This detailed entity annotation schema provided a robust foundation for subsequent tasks, enabling a nuanced understanding of text data.

The RDC task, introduced in the program's second phase, was pivotal in identifying and characterizing the relations between entities. This addition significantly expanded the scope of the ACE dataset, incorporating a variety of relations such as physical, social/personal, employment/membership, and more. The emphasis on capturing relations supported by textual evidence versus those inferred contextually introduced a layer of complexity, pushing forward the capabilities in relation extraction technologies.

With the introduction of EDC in ACE Phase 3, the program took on the new challenge of identifying and categorizing events in which entities participate. This expanded the dataset's utility by providing insights into interactions, movements, transfers, creations, and destructions depicted in text, along with event arguments and attributes based on type-specific templates. Later phases further enriched the dataset with additional event types and characterized relations between events, offering an even more comprehensive resource for NER and RE tasks.

The significance of the ACE dataset to NER and RE tasks lies in its comprehensive coverage of entities, relations, and events, making it a cornerstone in the development of technologies for information extraction. By providing a structured framework for annotating and understanding complex language data, the ACE dataset has been instrumental in advancing research and applications in NER and RE, enabling more sophisticated and nuanced language processing capabilities.

2.2.2 The SciERC dataset

The SciERC dataset, meticulously crafted from the domain of scientific research papers, particularly in the field of Artificial Intelligence (AI), represents a significant advancement in the realm of NER and RE tasks[8]. Developed by the Allen Institute for AI, SciERC's primary objective is to facilitate the extraction of scientific entities, their relationships, and events from AI research literature, thereby enabling a deeper understanding and structuring of scientific knowledge. This dataset emerges from the recognition of the unique challenges presented by scientific texts, which include domain-specific terminology, complex entity relations, and the nuanced depiction of scientific events and processes.

SciERC is distinguished by its focus on scientific texts, comprising 500 abstracts from AI conference proceedings, annotated for entities, relations, and coreference clusters. Entities within SciERC are categorized into specific types such as tasks, methods, metrics, materials, and others relevant to scientific discourse. This categorization facilitates a granular understanding of the scientific narrative, allowing for the extraction of nuanced information regarding the methodologies, tools, and outcomes prevalent within AI research. Furthermore, the dataset annotates relations between these entities, providing insights into the interdependencies and interactions that define scientific innovation. Such detailed annotation makes SciERC an invaluable resource for developing NER and RE models tailored to the scientific domain.

The significance of the SciERC dataset to NER and RE tasks extends beyond its domain-specific focus. By offering a structured framework for analyzing scientific texts, SciERC enables the development of models capable of navigating the complexities inherent in scientific literature. These models are not only instrumental in extracting information from research papers but also in facilitating the synthesis of scientific knowledge, contributing to meta-analyses, systematic

reviews, and the construction of scientific knowledge graphs. In this way, SciERC supports the broader objective of making scientific knowledge more accessible and understandable, both to machines and to humans.

The connection between the SciERC dataset and our task is particularly poignant. Given the thesis's focus on extracting structured information from research papers within a specific scientific domain, SciERC provides a relevant model for addressing similar challenges in our research. The methodologies and insights gained from working with the SciERC dataset can inform the development of specialized NER and RE models for fields other than AI, enabling the extraction of entities and relations specific to those fields. Moreover, the success of models trained on SciERC underscores the potential for applying advanced NER and RE techniques to a wide range of scientific disciplines, thereby enhancing the accessibility and interoperability of scientific knowledge across domains.

The SciERC dataset represents a pivotal resource for advancing NER and RE tasks within the scientific domain. Its focus on AI research literature not only addresses the specific challenges of scientific text analysis but also offers a blueprint for extending these capabilities to other scientific disciplines. By enabling the development of models that can accurately identify and relate entities within scientific texts, SciERC contributes to the broader goal of structuring scientific knowledge, making it more navigable and comprehensible for both academic and practical purposes.

2.2.3 Other datasets

In our research, we have come to study other datasets that must be mentioned. Those datasets were used as a sanity check for our work. We relied on those datasets to prove that our models could be generalized and used for NER and RE tasks in other domains.

The New York Times Relation Extraction Dataset

The New York Times Relation Extraction Dataset (NYT dataset) is a prominent resource for RE, offering a comprehensive collection of news articles for the development and testing of RE models. Originating from a collaboration between the New York Times and Google, this dataset encompasses a vast array of articles published by the New York Times, annotated with both entities and

the relations between them[9]. The primary objective of the NYT dataset is to support the extraction of semantic relationships within text, facilitating a deeper understanding of the interconnectedness of entities as reported in journalistic content.

The dataset is characterized by its diverse coverage of topics, including politics, sports, culture, and more, reflecting the wide-ranging nature of news reporting. This diversity presents unique challenges and opportunities for RE, requiring models to adapt to various contexts and entity types. Each article within the dataset is annotated with detailed information about entities and the specific relations that link them, providing a rich ground for training sophisticated RE models capable of identifying and classifying a wide range of relation types.

The significance of the NYT dataset to the RE task lies in its real-world applicability and the complexity of its textual content. Working with this dataset enables researchers to hone RE models on text that encapsulates a broad spectrum of human activity and knowledge, mirroring the complexity and nuance of natural language used in daily news cycles. Additionally, the NYT dataset serves as a benchmark for evaluating the performance of RE models, offering a standard against which to measure progress in the field.

The NYT dataset exemplifies the application of RE techniques to general-domain texts, contrasting with the specialized domain of marine biology research papers. The exploration of this dataset highlights the adaptability of RE methodologies across different textual domains, underscoring the potential for leveraging insights gained from working with the NYT dataset to enhance RE approaches tailored to scientific literature. This cross-domain exploration illustrates the broad applicability of RE technologies and the importance of diverse datasets in advancing the field.

Text Analysis Conference Relation Extraction Dataset (TACRED)

The Text Analysis Conference (TAC) Relation Extraction Dataset is a crucial dataset in the domain of Relation Extraction, developed under the auspices of the TAC Knowledge Base Population (KBP) evaluations. Managed by the National National Institute of Standards and Technology (NIST), the TAC KBP evaluations are designed to foster research and development in the field of information extraction, with a focus on building comprehensive knowledge bases

from unstructured text[10]. The TAC RE dataset specifically aims to advance the state of RE technology by providing a set of documents annotated with entities and their relations, serving as both a training and evaluation resource for RE systems.

The dataset encompasses a diverse collection of texts sourced from newswire and web texts, including a wide range of topics and entity types. Entities within the dataset are meticulously annotated, and the dataset identifies various types of semantic relations that occur between these entities, such as affiliation, personal/social relationships, and organizational roles, among others. This rich annotation scheme allows for the detailed examination and modeling of complex relationships within natural language, making the TAC RE dataset an invaluable resource for researchers and developers working on advanced RE systems.

The significance of the TAC RE dataset extends beyond its comprehensive annotations; it also serves as a benchmark for evaluating the performance of RE systems in a competitive and collaborative environment. Through the annual TAC KBP evaluations, participating systems are assessed on their ability to accurately identify and characterize relations between entities, fostering innovation and progress in the field. The dataset not only facilitates the development of more sophisticated and accurate RE models but also promotes the exploration of new methodologies and approaches in knowledge base population.

Including this dataset in our research underscores its role in pushing the boundaries of RE technology. While the TAC dataset focuses on a general domain, the methodologies and insights derived from working with this dataset are directly applicable to the specialized domain of marine biology research papers. The challenges and solutions encountered in the TAC RE dataset provide a valuable perspective on the adaptation of RE techniques to domain-specific needs, demonstrating how RE technologies can be leveraged to extract structured information from diverse sources of text.

2.3 Exisitng NER and RE Models

Many models have been proposed for tackling NER and RE tasks. And in recent years there's been an emphasis on joint models. Join models are designed to perform join extraction of entities and relations[1] at the same time. We can group existing joint models into two categories: structured prediction and

multi-task learning.

2.3.1 Structured prediction models

Structured prediction approaches cast the two tasks into one unified framework, although it can be formulated in various ways.
Li and Ji[11] proposed an action-based system that identifies new entities as well as links to previous entities, Zhang et al.[12]; A novel and impactful methodology for the incremental joint extraction of entity mentions and relations. Their approach diverges from traditional methods by employing a structured perceptron with beam-search, moving away from token-based tagging to a segment-based decoder inspired by semi-Markov chains. This shift allows for the utilization of global features as soft constraints, effectively capturing the interdependencies between entities and relations. Their research, conducted on the ACE dataset, demonstrates significant advancements over existing pipelined approaches. By formulating the problem as one of structured prediction, their model adeptly captures the linguistic and logical nuances inherent in complex textual relationships, thereby addressing the limitations of sequential classification steps that fail to model long-distance and cross-task dependencies. The introduction of novel global features based on soft constraints over the entire output graph structure marks a significant contribution to the field, showcasing the potential for improved accuracy and efficiency in the extraction tasks. Li and Ji's work stands as a pivotal reference in the exploration of joint models and global features for enhancing entity and relation extraction, offering valuable insights and methodologies that could be adapted and extended within the context of our research.

Wang and Lu[13] adopt a table-filling approach as proposed in (Miwa and Sasaki[14]); This distinct approach departs from traditional single-encoder methods. Their method introduces two specialized encoders: a table encoder and a sequence encoder, designed to synergize in the representation learning process for NER and RE. This allows each encoder to focus on the unique aspects of its task—capturing task-specific information effectively—while benefiting from the interaction between the two to enhance overall performance. They leverage multi-dimensional recurrent neural networks to better utilize the structural information within the table representation, addressing a common limitation in

existing methods that often overlook or underutilize such information. Furthermore, they exploit the pairwise self-attention weights from pre-trained models like BERT[15] to enrich their model's understanding of word-word interactions, a strategy not previously employed for table representations in this context. Their experiments across several standard datasets show significant improvements over existing approaches, particularly highlighting the advantage of dual encoders over traditional single-encoder frameworks. This work not only sets new state-of-the-art performance benchmarks but also opens up new avenues for leveraging the inherent structure in linguistic data for information extraction tasks.

Katiyar and Cardie[16] and Zheng et al.[17] introduced an approach based on sequence-tagging for the joint extraction of entity mentions and relations, each contributing novel methodologies to the domain of information extraction. Katiyar and Cardie introduce an attention-based recurrent neural network model that leverages Long Short-Term Memory (LSTM) networks to extract semantic relations between entity mentions without relying on dependency trees. Their model is distinct for its direct addressing of the relation extraction task by integrating attention mechanisms with LSTMs, enabling the model to focus on relevant parts of the text to better identify relationships between entities, even when they are not adjacent. This approach sidesteps the need for dependency tree information, making it more broadly applicable, especially for languages or domains where dependency parsing might be less accurate or entirely unavailable. Their experiments on the ACE dataset demonstrate significant improvements over previously established feature-based joint models, highlighting the efficacy of their methodology in enhancing the accuracy of both entity and relation extraction tasks.

Zheng et al. propose a different take on sequence tagging by introducing a novel tagging scheme that converts the joint task of entity and relation extraction into a single tagging problem. This simplifies the traditionally complex process of first identifying entities and then classifying relations between them. Their end-to-end model, also based on LSTM networks, directly extracts entities and their relations without the need for separate entity recognition and relation classification stages. By treating the problem as a tagging issue, they manage to avoid the error propagation and complexity associated with pipelined and feature-based methods. Their approach not only demonstrates superior per-

formance on a public dataset produced by distant supervision methods but also underscores the potential of tagging-based methods in streamlining the extraction process and improving result accuracy.

These contributions represent significant advancements in the field of information extraction, particularly in the context of NER and RE. They offer insights into the potential of neural network architectures and tagging schemes to simplify and enhance the joint extraction of entities and relations, paving the way for more efficient and accurate extraction methodologies suitable for a wide range of applications.
Sun et al.[18] and Fu et al.[19] used a graph-based method to predict entity and relation types, offering significant advancements in joint entity and relation extraction tasks. Sun et al. introduced a novel Graph Convolutional Network (GCN) approach that operates on an entity-relation bipartite graph, designed to perform joint inference on entity types and relation types within a unified framework. This method significantly outperformed existing joint models in entity performance while maintaining competitive relation performance on the ACE05 dataset. The key to their approach was the introduction of a binary relation classification task that allowed for more efficient and interpretable use of the entity-relation bipartite graph structure.

Fu et al. presented GraphRel, an end-to-end relation extraction model employing GCNs to jointly learn named entities and relations. By considering both the interaction between named entities and relations and the implicit features among all word pairs in the text, GraphRel demonstrated substantial improvements in predicting overlapping relations compared to previous sequential approaches. Their graph-based strategy, which utilized both linear and dependency structures, alongside a complete word graph to extract features, resulted in high precision and a significant increase in recall, setting new state-of-the-art benchmarks for relation extraction on public datasets like NYT and WebNLG.

These contributions reflect a deeper understanding of how entities and relations interconnect within text, underscoring the potential of graph-based models to capture complex relationships more effectively than traditional methods. Through the integration of GCNs and strategic graph construction, both approaches highlight the evolving landscape of natural language processing, where the interconnectedness of textual elements is increasingly recognized and leveraged for more nuanced and accurate information extraction.

and, Li et al[20] project the task onto a multi-turn question answering problem, transforming the entity and relation extraction process into an innovative QA framework. This paradigm shift offers several advantages: it encodes specific class information for the desired entity or relation through the formulation of questions, naturally incorporates joint modeling of entities and relations, and leverages advanced Machine Reading Comprehension (MRC) models. Their approach not only significantly outperforms existing models on benchmark datasets like ACE and CoNLL04 but also establishes new state-of-the-art results, highlighting its effectiveness in accurately identifying structured information from text. Moreover, Li et al. introduce a complex dataset, RESUME, requiring multi-step reasoning for entity dependency construction, further demonstrating the model's capability in handling intricate entity-relation mappings. This multi-turn QA framework marks a substantial advance in entity-relation extraction, showing promise for more nuanced and accurate information extraction from unstructured data.
All of these approaches need to tackle a global optimization problem and perform joint decoding at inference time, using beam search or reinforcement learning.

In general, structured prediction models are challenged by the complexity in modeling interdependencies. These models attempt to capture the complex interdependencies between entities and relations within a single framework, which can be computationally intensive and challenging to optimize. In addition, they also have to deal with the complexity of joint decoding. Performing joint decoding at inference time, such as using beam search or reinforcement learning, adds to the computational overhead and complexity.

2.3.2 Multi-task learning models

This family of models essentially builds two separate models for entity recognition and relation extraction and optimizes them together through parameter sharing.
Miwa and Bansal[21] propose to use a sequence tagging model for entity prediction and a tree-based LSTM model for relation extraction. The two models share one LSTM layer for contextualized word representations and they find sharing parameters improves performance (slightly) for both models. Their innovative approach captures both word sequence and dependency tree sub-

structure information, integrating bidirectional tree-structured LSTM-RNNs on top of bidirectional sequential LSTM-RNNs. This allows for a single model to jointly represent entities and relations with shared parameters, improving over the state-of-the-art feature-based models on end-to-end relation extraction tasks. By encouraging the detection of entities during training and utilizing entity information in relation extraction through entity pretraining and scheduled sampling, Miwa and Bansal's model demonstrates substantial error reductions in F1-score on both ACE2005 and ACE2004 datasets, setting new benchmarks for the field. Their work underscores the importance of combining linear sequence and tree structure representations for capturing the nuances of entity and relation extraction in text. The approach of Bekoulis et al.[22] is similar except that they model relation classification as a multi-label head selection problem. Note that these approaches still perform pipelined decoding: entities are first extracted and the relation model is applied on the predicted entities. In their work on adversarial training for multi-context joint entity and relation extraction, Bekoulis et al. extend a baseline joint model that tackles NER and RE simultaneously, by introducing Adversarial Training (AT) as a regularization method. This technique enhances the model's robustness by incorporating small perturbations in the training data, thereby improving the state-of-the-art effectiveness across several datasets and languages. Their model successfully addresses the complexities of extracting multiple relations per entity by modeling relation extraction in a multi-label setting, allowing for a more nuanced understanding of the text. Additionally, their innovative use of AT demonstrates a significant improvement in the joint extraction task's effectiveness, showcasing the potential of adversarial examples in NLP to refine and strengthen model performance.

Dynamic Graph Interaction Extraction (DYGIE) and DYGIE++ (Luan et al. [23]; Wadden et al. [24]), build on recent span-based models for coreference resolution (Lee et al. [25]) and semantic role labeling (He et al. [26]). The key idea of their approaches is to learn shared span representations between the two tasks and update span representations through dynamic graph propagation layers. DYGIE++ extends upon these concepts by incorporating event extraction into its multi-task framework, utilizing both local (within-sentence) and global (cross-sentence) context to enumerate, refine, and score text spans. The system dynamically constructs graphs of spans, with edges representing task-specific relations, allowing for efficient global context modeling. This is achieved by

refining initial contextualized embeddings, such as those from Bidirectional Encoder Representations from Transformers (BERT), with task-specific message updates propagated across the span graph.

The DYGIE++ framework demonstrates its effectiveness by achieving state-of-the-art results across several information extraction tasks and datasets, showcasing its capability to handle complex interdependencies among entities, relations, and events. The integration of BERT encodings enables the model to capture significant contextual relationships, including those extending beyond single sentences. Additionally, dynamic span graph updates further enhance the model's ability to incorporate cross-sentence dependencies, which is particularly beneficial for tasks in specialized domains. For example, leveraging predicted coreference links through graph propagation can help disambiguate challenging entity mentions by providing additional contextual clues.

A comprehensive evaluation of the DYGIE++ framework across different datasets reveals that its general span-based approach produces significant improvements in entity recognition, relation extraction, and event extraction tasks. The framework benefits from both types of contextualization methods—BERT encodings for capturing immediate and adjacent-sentence context, and message passing updates for modeling long-range cross-sentence dependencies. These findings underscore the importance of effectively integrating both local and global contextual information in a unified architecture to enhance performance on a range of information extraction tasks, making DYGIE++ a powerful tool for advancing research in this area.

A more recent work Lin et al.[27] further extends DYGIE++ by incorporating global features based on cross-substask and cross-instance constraints. They propose a joint neural framework named ONEIE, which aims to extract the globally optimal Information Extraction (IE) result as a graph from an input sentence. This framework performs IE in four stages: encoding the given sentence as contextualized word representations; identifying entity mentions and event triggers as nodes; computing label scores for all nodes and their pairwise links using local classifiers; and finally, searching for the globally optimal graph with a beam decoder. At the decoding stage, they introduce global features to capture the intricate cross-subtask and cross-instance interactions. Their experimental results demonstrate that adding these global features significantly improves the performance of their model, achieving new state-of-the-art results on all subtasks. Unlike previous models that use separate local task-specific classifiers in

their final layer without explicitly modeling the dependencies among tasks and instances, ONEIE extracts a unified graph representation of the input sentence, effectively capturing and leveraging the interdependencies among different IE components. This advancement underscores the importance of considering the holistic context of information in IE tasks, marking a significant step forward in the development of more integrated and contextually aware IE systems.
Zhong et al. [1] introduced the Princeton University Relation Extraction system (PURE), a similar approach. However, it is much simpler and performs better. Their model challenges the longstanding belief in the superiority of complex joint models for entity and relation extraction tasks. Through their research, they illuminate the effectiveness of a straightforward pipelined approach that employs two independent encoders for entity recognition and relation extraction, both built upon deep pre-trained language models. Their method deviates from the common practice of intricate joint modeling, advocating instead for simplicity and directness in treating the tasks sequentially. This simplicity, coupled with meticulous analyses on standard benchmarks like ACE04, ACE05, and SciERC, not only sets new state-of-the-art performances with absolute improvements in relation to F1 scores but also demonstrates the critical importance of learning distinct contextual representations for entities and relations. Furthermore, their investigation into incorporating entity information early in the relation model underscores the potential of a more focused approach to enhancing performance.

Their work significantly contributes to the discourse on the efficiency of information extraction models, showing that a model's complexity does not necessarily equate to its effectiveness. By simplifying the process into two distinct phases and ensuring each phase is optimized for its specific task, they reveal an often-overlooked aspect of model design: the power of specialization and focused optimization. Their findings suggest that the interactions between entities and relations, previously believed to be best captured jointly, can be effectively understood through a well-structured sequential approach. This revelation opens new avenues for future research in information extraction, particularly in exploring how different tasks within this domain can be optimized individually for better overall performance.

Moreover, the authors explored the utility of pre-trained language models as a foundation for both encoders bringing to light the substantial impact of these models in extracting meaningful insights from text. By leveraging such

powerful models, they manage to streamline the extraction process and ensure that their approach remains flexible and robust across various datasets. This adaptability, combined with the method's simplicity, marks a significant step forward in information extraction research. It prompts a reevaluation of current methodologies and suggests that the field might benefit from a shift towards simpler, more focused models that capitalize on the advancements in language modeling and representation learning.

However impressive, their model still faces several challenges. These include the potential for reduced effectiveness on rare or unseen entities, increased computational demands due to its complexity, reliance on accurate entity type identification, and difficulties in handling ambiguous contexts or adapting to various languages and domains. Moreover, the model risks overfitting to training data and may present challenges in interpretability, making it harder to understand how it makes decisions. Addressing these issues is essential for optimizing the model's performance and applicability across diverse datasets and settings.

3

Methodology

In this chapter, we provide a formal definition of the problem of joint entity and relation extraction in section [3.1]. Then, in section [3.2], we describe our approach in detail. First explaining the entity model in section [3.2.1], and then describing the relation model in section [3.2.2]. Finally, we explain the training and inference processes in section [3.2.3]

3.1 Problem Definition

Given X an input sentence consisting of n tokens $x_1, x_2, ..., x_n$. Let $S = s_1, s_2, ..., s_m$ be all the possible spans in X of up to length L and $START(i)$ and $END(i)$ denote start and end indices of s_i. The problem can be decomposed into two sub-tasks:

Named entity recognition Let E denote a set of pre-defined entity types. The named entity recognition task is, for each span $s_i \in S$, to predict an entity type

$$y_e(s_i) \in E$$

or, span s_i is not an entity:

$$y_e(s_i) = \epsilon$$

The output of the task is

$$Y_e = (s_i, e) : s_i \in S, e \in E$$

.

Relation extraction Let R denote a set of predefined relation types. The task is, for every pair of spans $si \in S$, $sj \in S$, to predict a relation type

$$y_r(s_i, s_j) \in R$$

, or there is no relation between them:

$$y_r(s_i, s_j) = \epsilon$$

. The output of the task is

$$Y_r = (s_i, s_j, r) : s_i, s_j \in S, r \in R$$

.

3.2 Our Approach

We based our approach on the state of the art proposed by Zhong et al. [1]. The simplicity of their model and its performance make it a prime candidate for NER and RE on new datasets. Therefore we first reproduced their results on the SciERC dataset. Then, we prove that their model can be generalized to other datasets. In the next chapter, we will dive deeper into our experiments and demonstrate how we trained and evaluated their model on new NER and RE datasets. Namely the NYT [2.2.3.1] and TACRED[2.2.3.2] datasets. The goal is to have a framework that, given any NER and RE dataset, can be easily used to train NER and RE models.

3.2.1 Entity model

The entity model is inspired by previous research (Lee et al. [25]; Luan et al. [23]; Wadden et al. [24]). It begins by using a pre-trained language model, such as BERT, to understand the context of each word in a sentence. For any given segment of text, known as a 'span', we create a context representation x_t for each input token x_t. This is done by combining the context of the span's first and last words with additional features that capture the span's length. We then

use this combined information to calculate the likelihood of each possible entity type that this span could represent.
Formally, Let S be an input sentence, and s_i a span of S $s_i \in S$, we can define the span representation $h_e(s_i)$ as[1]:

$$h_e(s_i) = [x_{START(i)}; x_{END(i)}; \phi(s_i)]$$

, where $\phi(s_i) \in R_F^d$ is a representation of the span width features. Finally, $h_e(s_i)$ becomes the input to a Feedforward Netwrok (FFN) that will predict the entity type. Or, more precisely, its probability distribution:

$$e \in E \cup \epsilon : P_e(e|s_i)$$

.

3.2.2 Relation Model

The relation model function is to take two spans s_i , s_j, which are the 'subject' and 'object' as input and output a relation between them. Or output ϵ if there's no relation. Most works we examined (Luan et al. [23]; Wadden et al. [24]) use the same span representations $h_e(s_i)$, $h_e(s_j)$ in the relation model to predict the relation between s_i and s_j . Hover, Zhong et al. [1] suggest that while these representations can understand the context surrounding each entity independently, they might not effectively identify the connections or relationships between pairs of text segments. They also point out that using the same contextual information for different pairs of text segments might not always be the best approach. For example, the phrase "is a" is important for recognizing the relationship between MORPA and PARSER in Figure 1, but does not help in understanding the connection between MORPA and TEXT-TO-SPEECH.

Instead, Zhong et al. [1] propose a relation model that looks at each pair of spans separately and adds specific markers in the initial processing stage. These markers indicate which span is the subject and which is the object, as well as their types, to improve the model's understanding. Formally, Let X be an input sentence and s_i , s_j be a pair of subject-object spans and $ei, ej \in E \cup \epsilon$ are their types respectively. Then we define text markers as $\langle S : e_i \rangle$, $\langle /S : e_i \rangle$, $\langle O : e_j \rangle$, and $\langle /O : e_j \rangle$, and embedded them into X before and after s_i and s_j (Figure 1

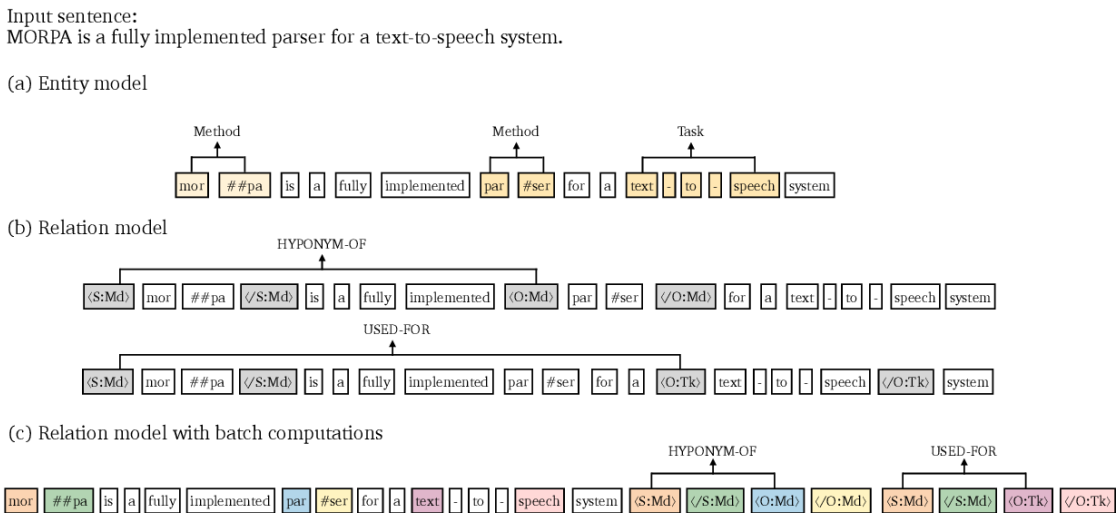

Figure 3.1: An input sentence from the SciERC dataset. Luan et al. [23]

(b)). Let $\widehat{X}$ be the new sequence after inserting the markers:

$$\widehat{X} = ...\langle S : e_i\rangle, x_{START(i)}, ..., x_{END(i)}, \langle /S : e_i\rangle ...\langle /O : e_j\rangle, x_{START(j)}, ..., x_{END(j)}\langle /O : e_j\rangle$$

Next, we use another pre-trained encoder on $\widehat{X}$ and we refer to its output representations with by $\widehat{x}_t$. We combine the outputs from their starting points to understand the relationship between the two spans. This gives us a combined representation:

$$h_r(s_i, s_j) = [\widehat{x}_{\widehat{START(i)}}; \widehat{x}_{\widehat{START(j)}}]$$

where $\widehat{START(i)}$ and $\widehat{START(j)}$ are the indices of $\langle S : e_i\rangle$ and $\langle O : e_j\rangle$ in $\widehat{X}$. Finally, $h_r(s_i, s_j)$ will be the input to an FFN that will predict the relation between s_i and s_j:

$$r \in R \cup \epsilon : P_r(r|s_i, s_j)$$

.

The approach of using special markers to identify subjects and objects in a text is not particularly novel and has been explored before in classification studies (Zhang et al.[28]; Soares et al.[29]; Peters et al.[30]). However, these studies usually focus on classifying the relationship between one pair of subjects and objects within a sentence Zhang et al.[31], such as in the TACRED dataset[2.2.3.2]. The effectiveness of this method has not been fully tested in the end-to-end setting like the Zhong et al. [1] hope to classify the relations amongst more entity mentions. They saw a significant improvement in their

solution, strengthening the idea that different context-aware representations are invaluable for understanding the relations between entities pairs in one example.. Furthermore, Zhang et al.[28]; Soares et al.[29] used untyped only markers (e.g., $\langle S:\rangle$, $\langle /S:\rangle$) and previous end-to-end models (e.g., (Wadden et al.[24])) only inject the entity type information into the relation model through auxiliary losses. Zhong et al. [1] found that injecting type information at the input layer is very helpful in distinguishing entity types — for example, whether "Disney" refers to a person or an organization— before trying to understand the relations.

3.2.3 Training and Inference

We adapt two pre-trained language models by fine-tuning them with task-specific loss functions, employing cross-entropy loss[1], for both the entity and relation extraction models. This equation penalizes the model more heavily when its predicted probability for the true entity type is lower, encouraging the model to correctly recognize entity types.

$$\mathcal{L}_e = -\sum_{s_i \in S} \log P_e(e_i^* | s_i)$$

Where $\mathcal{L}_e$ represents the cross-entropy loss function for the entity model. It is calculated by summing the negative log probabilities of the true (gold) entity types (e_i^*) for all spans (s_i) in the dataset (S). The probability $P_e(e_i^*|s_i)$ reflects how likely it is that the span s_i corresponds to its gold entity type e_i^* according to the model.

$$\mathcal{L}_r = -\sum_{s_i, s_j \in S, s_i \neq s_j} \log P_r(r_{i,j}^* | s_i, s_j)$$

Where $\mathcal{L}_r$ represents the cross-entropy loss function for the relation model. Similar to $\mathcal{L}_e$, this loss function sums the negative log probabilities that the model assigns to the true (gold) relation types ($r_{i,j}^*$) between pairs of spans (s_i, s_j) in the dataset (S), where $s_i \neq s_j$. The probability $P_r(r_{i,j}^*|s_i, s_j)$ indicates the model's confidence that the correct relation between the spans $s_i$and s_j is $r_{i,j}^*$.

where e_i^* represents the gold entity type of s_i and $r_{i,j}^*$ represents the gold relation type of span pair s_i, s_j in the training data. For training the relation model, we only consider the gold entities $S_G \subset S$ in the training set and use the gold entity labels as the input of the relation model. During inference, we first

predict the entities by taking $y_e(s_i) = argmax_{e \in \varepsilon \cup \{\epsilon\}} P_e(e|s_i)$. Denote $S_{pred} = s_i : y_e(s_i) \neq \epsilon$, we enumerate all the spans $s_i, s_j \in S_{pred}$ and use $y_e(s_i), y_e(s_j)$ to construct the input for the relation model $P_r(r|s_i, s_j)$.

In training the relation model, we focus exclusively on the spans marked as entities according to the gold standard S_G, which is a subset of all spans S in the dataset ($S_G \subset S$). This allows the model to learn from the most relevant examples, using the gold entity labels to understand relationships within the text.

During the inference phase, the model predicts entities by determining the most likely entity type for each span s_i. That's done by taking

$$y_e(s_i) = argmax_{e \in \varepsilon \cup \{\epsilon\}} P_e(e|s_i)$$

, where ε includes all possible entity types and ϵ indicates a non-entity. Spans not identified as entities are filtered out, creating a set $S_{pred} = s_i : y_e(s_i) \neq \epsilon$ of spans predicted to be entities.

With S_{pred} established, the model then examines every possible pair of spans within it to predict their interrelations. For each span pair s_i, s_j, it employs the predicted entity types as inputs to the relation model $P_r(r|s_i, s_j)$. This step utilizes the insights gained from entity prediction to enhance the accuracy of relation extraction, aiming to comprehensively map out the network of relationships among identified entities in the text. This process underscores the model's integrated approach, leveraging entity predictions to inform and refine relation extraction, thus creating a cohesive understanding of the textual content.

4

Experiments

5

Results

6

Conclusions and Future Works

References

[1] Zhong, Z. and Chen, D. "A Frustratingly Easy Approach for Joint Entity and Relation Extraction". In: *ArXiv* abs/2010.12812 (2020). URL: https://api.semanticscholar.org/CorpusID:232320859.

[2] Goyal, A., Gupta, V., and Kumar, M. "Recent Named Entity Recognition and Classification techniques: A systematic review". In: *Computer Science Review* 29 (Aug. 2018), pp. 21–43.

[3] Fang, Z. et al. "TEBNER: Domain Specific Named Entity Recognition with Type Expanded Boundary-aware Network". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Moens, M.-F. et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 198–207. URL: https://aclanthology.org/2021.emnlp-main.18.

[4] Bhandari, N. et al. "Resolving Ambiguities in Named Entity Recognition Using Machine Learning". In: *2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)*. 2017, pp. 159–163.

[5] Tsai, C.-T., Mayhew, S., and Roth, D. "Cross-Lingual Named Entity Recognition via Wikification". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Ed. by Riezler, S. and Goldberg, Y. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 219–228. URL: https://aclanthology.org/K16-1022.

[6] Le, P. and Titov, I. *Improving Entity Linking by Modeling Latent Relations between Mentions*. 2018. arXiv: 1804.10637 [cs.CL].

[7] Linguistic Data Consortium, T.T.o.t.U.o.P. *ACE*. https://www.ldc.upenn.edu/collaborations/past-projects/ace.

[8] Luan, Y. et al. "Multi-Task Identification of Entities, Relations, and Coreferencefor Scientific Knowledge Graph Construction". In: *Proc. Conf. Empirical Methods Natural Language Process. (EMNLP)*. 2018.

[9] Tripathi, S. *New York Times Relation Extraction Dataset*. https://www.kaggle.com/datasets/daishinkan002/new-york-times-relation-extraction-dataset/data.

[10] Zhong Victor, e.a. *TAC Relation Extraction Dataset LDC2018T24*. https://catalog.ldc.upenn.edu/LDC2018T24. Dec. 2018.

[11] Li, Q. and Ji, H. "Incremental Joint Extraction of Entity Mentions and Relations". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Toutanova, K. and Wu, H. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 402–412. URL: https://aclanthology.org/P14-1038.

[12] Zhang, M., Zhang, Y., and Fu, G. "End-to-End Neural Relation Extraction with Global Optimization". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Palmer, M., Hwa, R., and Riedel, S. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 1730–1740. URL: https://aclanthology.org/D17-1182.

[13] Wang, J. and Lu, W. "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)*. Ed. by Webber, B. et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 1706–1721. URL: https://aclanthology.org/2020.emnlp-main.133.

[14] Miwa, M. and Sasaki, Y. "Modeling Joint Entity and Relation Extraction with Table Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Moschitti, A., Pang, B., and Daelemans, W. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1858–1869. URL: https://aclanthology.org/D14-1200.

[15] Devlin, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[16] Katiyar, A. and Cardie, C. "Going out on a limb: Joint Extraction of Entity Mentions and Relations without Dependency Trees". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Barzilay, R. and Kan, M.-Y. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 917–928. URL: https://aclanthology.org/P17-1085.

[17] Zheng, S. et al. "Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Barzilay, R. and Kan, M.-Y. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1227–1236. URL: https://aclanthology.org/P17-1113.

[18] Sun, C. et al. "Joint Type Inference on Entities and Relations via Graph Convolutional Networks". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Korhonen, A., Traum, D., and Màrquez, L. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1361–1370. URL: https://aclanthology.org/P19-1131.

[19] Fu, T.-J., Li, P.-H., and Ma, W.-Y. "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*. Ed. by Korhonen, A., Traum, D., and Màrquez, L. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1409–1418. URL: https://aclanthology.org/P19-1136.

[20] Li, X. et al. "Entity-Relation Extraction as Multi-Turn Question Answering". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Korhonen, A., Traum, D., and Màrquez, L. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1340–1350. URL: https://aclanthology.org/P19-1129.

[21] Miwa, M. and Bansal, M. "End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Erk, K. and Smith, N.A. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1105–1116. URL: https://aclanthology.org/P16-1105.

[22] Bekoulis, G. et al. "Adversarial training for multi-context joint entity and relation extraction". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Riloff, E. et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2830–2836. URL: https://aclanthology.org/D18-1307.

[23] Luan, Y. et al. "A general framework for information extraction using dynamic span graphs". 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)*. Ed. by Burstein, J., Doran, C., and Solorio, T. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3036–3046. URL: https://aclanthology.org/N19-1308.

[24] Wadden, D. et al. "Entity, Relation, and Event Extraction with Contextualized Span Representations". In: *ArXiv* abs/1909.03546 (2019). URL: https://api.semanticscholar.org/CorpusID:202539496.

[25] Lee, K. et al. "End-to-end Neural Coreference Resolution". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Palmer, M., Hwa, R., and Riedel, S. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 188–197. URL: https://aclanthology.org/D17-1018.

[26] He, L. et al. "Jointly Predicting Predicates and Arguments in Neural Semantic Role Labeling". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Gurevych, I. and Miyao, Y. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 364–369. URL: https://aclanthology.org/P18-2058.

[27] Lin, Y. et al. "A Joint Neural Model for Information Extraction with Global Features". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Jurafsky, D. et al. Online: Association for Computational Linguistics, July 2020, pp. 7999–8009. URL: https://aclanthology.org/2020.acl-main.713.

[28] Zhang, Z. et al. "ERNIE: Enhanced Language Representation with Informative Entities". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Korhonen, A., Traum, D., and

Màrquez, L. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1441–1451. URL: https://aclanthology.org/P19-1139.

[29] Baldini Soares, L. et al. "Matching the Blanks: Distributional Similarity for Relation Learning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Korhonen, A., Traum, D., and Màrquez, L. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2895–2905. URL: https://aclanthology.org/P19-1279.

[30] Peters, M.E. et al. *Knowledge Enhanced Contextual Word Representations*. 2019. arXiv: 1909.04164 [cs.CL].

[31] Zhang, Y. et al. "Position-aware Attention and Supervised Data Improve Slot Filling". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Palmer, M., Hwa, R., and Riedel, S. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 35–45. URL: https://aclanthology.org/D17-1004.

Appendix