Knowledge Enabled Relation Extraction

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

MONIKA JAIN
PHD18021

COMPUTER SCIENCE AND ENGINEERING
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

19 June 2025

THESIS CERTIFICATE

This is to certify that the thesis titled **Knowledge Enabled Relation Extraction**, submitted by **Monika Jain**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. V. Raghava Mutharaju
Thesis Supervisor
Associate Professor, Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi.
Date: 19 June 2025
Place: New Delhi

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

DECLARATION

This is to be certified that the dissertation entitled **Knowledge Enabled Relation Extraction** being submitted by **Monika Jain** to the **Indraprastha Institute of Information Technology-Delhi**, for the award of degree of **Doctor of Philosophy**, is a bonafide work carried out by me. This research work has been carried out under the supervision of **Dr. V. Raghava Mutharaju**. The study pertains to this dissertation has not been submitted in part or in full, to any other University or Institution for the award of any other degree.

Monika Jain
PhD Student, PhD18021
Indraprastha Institute of Information Technology, Delhi.
Date: 19 June 2025
Place: New Delhi

ACKNOWLEDGEMENTS

I am greatly indebted to my PhD advisor, Prof. Raghava Mutharaju, for his mentorship and for pushing me to work to the best of my ability. He provided me with critical directions in technical matters, was ever willing to meet for discussions, and always had time to read drafts and comment on them. I appreciate his willingness to give me the independence I needed in my research and their immense role in improving my writing and presentation skills. I also wish to show my gratitude to the progress monitoring committee members, Dr. Rajiv Ratn Shah, IIIT-Delhi, and Dr. Vikram Goyal, IIIT-Delhi, for their valuable time reviewing my research and providing constructive feedback.

Special thanks to Dr. Kuldeep Singh, Cerence, Germany and Dr. Ramakanth Kavuluru, University of Kentucky, Lexington, USA, for being a fantastic mentor, for many hours of technical discussions that helped me grow as a researcher, and for an inspiring example of how to reconcile professional excellence.

My heartfelt thanks go to my beloved husband, Prateek, whose unwavering support, understanding, and sacrifices have been the cornerstone of my academic journey. His belief in me has been a constant source of strength, making this experience even more meaningful. To my child, Shubh, who spent much of his childhood in daycare while I worked on my PhD, you are my driving force and a reason to strive for more.

I am immensely grateful to my parents, siblings, and in-laws for their endless love, unwavering encouragement, and selfless sacrifices. Their unwavering support and faith in me have been pivotal to my academic achievement.

I owe my friend Gunjan Singh a debt of gratitude for the mental support and sense of peace she has given me from the very beginning. Special thanks to my friends Nidhi Goyal, Mayank Kharbanda, Apurv Dube, Aisha Aijaz, Avi Gupta, Saurabh Yadav and colleagues from the institute administration (IIIT-Delhi) for creating such a great working environment.

I express my sincere gratitude to the Infosys Centre for Artificial Intelligence (CAI) at IIIT-Delhi for their support throughout the duration of my PhD.

Lastly, I dedicate this work to my family, whose unwavering belief in me has been my greatest strength. This PhD is not merely my own but a shared success with all who have stood by me

throughout this journey.

ABSTRACT

Relation extraction is the task of extracting relationships between entities from text, where the text can be a part of a sentence, a document, or across multiple documents. This task has been popular for decades and is still of keen interest. Various techniques have been proposed to solve the relation extraction problem, among which the most popular are using distant supervision, deep learning-based models and reasoning-based models. However, these techniques rely on the knowledge between entities in a text and do not consider the background knowledge of the entities themselves, such as the entity type, synonyms and entity definitions. Predicting relations based on this knowledge is challenging due to the latent and unspecific contexts that introduce noise. To address these issues, we investigate mechanisms to incorporate background knowledge into the relation extraction task. We consider publicly available and relevant ontologies and knowledge graphs as sources of background knowledge. We propose three approaches named ReOnto, DocRE-CLip, and KXDocRE for relation extraction from text at three levels of granularity (sentence, document and across documents). These approaches embed knowledge in deep learning based models, and this has led to an improvement in their performance. We evaluate our approaches using domain-specific and general datasets. The results validate the utility of considering background knowledge for relation extraction.

KEYWORDS: Relation extraction, Domain knowledge, Distant supervision

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Relations between sets can be categorized based on the number of sets involved: unary, binary, and n-ary relations. A unary relation involves a single set and is simply a subset of that set. An n-ary relation involves n sets and is a subset of the cartesian product of these n sets. A binary relation involves two sets and is defined as a subset of the cartesian product $A \times B$, where the cartesian product $A \times B$ consists of all ordered pairs (a, b) such that $a \in A$ and $b \in B$.

In this work, we focus on binary relations instances, which are subsets of $A \times B$. In the context of relation extraction, the goal is to extract these binary relations from a given dataset or text, determining pairs of entities that are related in a specific way. In natural language processing, relation extraction (RE) tasks involves identifying and categorizing the relationships between entities within a text. It is crucial for structuring unstructured data, enabling more effective information retrieval, and enhancing the performance of various AI applications such as question-answering systems. For example, in the sentence provided in Fig. 1.1, the entities are `GE Aerospace`, `NASA` and `Aerospace technologies`. The relationship between `GE Aerospace` and `NASA` is `collaborates with`, and the relationship between `GE Aerospace` and `Aerospace technologies` is `works on`.

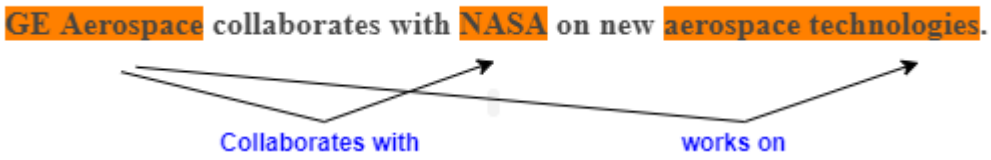

Figure 1.1: Simple example of relation extraction

Relation extraction enables machines to understand and interpret the complex web of relationships in human language. This process enables the conversion of textual data into a structured form, which is crucial for various applications. Its applications span across various fields.

Information Retrieval: Relation extraction can improve search algorithms by allowing them to understand the relationships between terms, leading to more accurate and relevant search results. This is particularly useful for legal documents, scientific research, and technical manuals [4].

Knowledge Graph Construction: By extracting relationships between entities, it's possible to build knowledge graphs that represent a network of interconnected information. These graphs

are essential for semantic search, recommendation systems, and AI applications that require a deep understanding of the relationships between different pieces of information [5]

Question Answering Systems: Relation extraction helps question answering systems understand the context and the specific relationships between entities mentioned in a question. It enables these systems to provide precise and accurate answers by focusing on the relevant connections between concepts [6].

Text Summarization: Understanding the key relationships between entities in a text can help generate concise summaries that capture the essence of the information. This is particularly useful for digesting large volumes of text quickly [7].

Sentiment Analysis: In sentiment analysis, relation extraction can identify the entities involved and the nature of their relationships, providing deeper insights into the sentiments expressed in the text. This is valuable for market analysis, brand monitoring, and social media analytics [8].

Compliance and Risk Management: In sectors like finance and healthcare, relation extraction can help identify non-compliance and potential risks by analyzing the relationships between various entities and actions within regulatory documents and communications [9].

Drug Discovery and Biomedical Research: In the biomedical field, relation extraction is used to identify relationships between genes, diseases, drugs, and other entities. This can accelerate research and discovery by uncovering hidden connections and insights from vast literature [10].

Machine Translation: Understanding the relationships between entities in a sentence can improve the accuracy of machine translation by ensuring that semantic relationships are preserved across languages [11].

1.1 Challenges

Although relation extraction has various applications, the process has several challenges and issues, such as ambiguity, contextual dependence, noise, lack of domain-specific knowledge, data sparsity, scalability, variability in language, and nested and overlapping relationships, among others. Let's look at some of the challenges of RE.

Ambiguity: Sometimes input text often contains ambiguous references to entities and relationships. For example, consider the sentence "Alice is close to Amazon". It is ambiguous whether "Amazon" refers to the rainforest or the company.

Contextual dependence: The meaning of a word or phrase can depend on the context in which it is used, which leads to difficulty in predicting the relation. For example: Depending on the

context,“pitch” can mean either “a presentation” or “to throw a ball” .

Domain specific knowledge: Extraction tasks often require domain-specific knowledge, which can be challenging to incorporate into extraction algorithms. For instance, a sample clinical note might contain a sentence such as: “Patient presents with SOB, hx of COPD, and was given 2L O2 via NC.” To accurately extract meaningful information from this sentence, an extraction algorithm needs to understand that: `SOB` stands for `shortness of breath`, `hx` stands for `history`, `COPD` stands for chronic obstructive pulmonary disease, 2L O2 via NC means 2 liters of oxygen via nasal cannula. Without domain-specific knowledge, an algorithm might struggle to interpret these abbreviations and terminologies correctly, leading to inaccurate or incomplete extraction of information. Incorporating domain-specific knowledge into the algorithm, such as a medical dictionary or ontology, can significantly improve the accuracy and reliability of the extraction process.

Data sparsity: Annotated training data for relation extraction tasks is often limited and expensive to create. This data sparsity can hinder the performance of machine learning models, especially for rare or complex relationships.

Scalability: Relation extraction systems must be scalable to process large volumes of text efficiently, especially in applications such as information retrieval or knowledge base construction.

Variability in language: Natural language is diverse and flexible, leading to variations in how relationships are expressed. Synonyms, abbreviations, and different ways of expressing the same relationship can make extraction challenging. For instance, given the sentences "John Doe is the CEO of TechCorp." and "JD serves as the Chief Executive Officer for TechCorp." knowing that CEO and Chief Executive Officer are synonyms allows us to understand that John Doe and JD are the same person.

Nested and overlapping relationships: Text can contain nested or overlapping relationships, where one entity is related to multiple other entities in different ways. Extracting such relationships accurately requires sophisticated algorithms. For example, consider the sentence “John, who is married to Jane, works at the same company as his friend Mark, who is also the brother of Jane”. Multiple relationships exist in this sentence that are both nested and overlapping such as “John is married to Jane”, “John works at the same company as Mark”, “Mark is the brother of Jane.”

Temporal Mismatch Between Document Content and Database: One significant challenge in relation extraction is the potential temporal mismatch between the content of a document and the information stored in a database. This issue arises when the document reflects outdated information, or the database has not been updated to reflect recent changes. For example, consider

a scenario where a document contains the sentence: “Rebecca works for XYZ Corporation”. If the knowledge base indicates that Rebecca has a “works for” relation with XYZ Corporation, the system might label the sentence accordingly. However, if Rebecca has recently left XYZ Corporation and this change is not reflected in the database, the extracted relation becomes inaccurate. This temporal mismatch can lead to incorrect labelling and inaccurate relation extraction, underscoring the importance of maintaining up-to-date databases and implementing mechanisms to verify the temporal consistency of extracted relations.

1.2 Motivation

In natural language understanding, extracting relations between entities from textual data has long been a pivotal challenge [12, 13]. As the volume and complexity of textual data continue to surge across diverse domains, the demand for advanced techniques to uncover hidden relationships between entities has intensified [14]. Furthermore, the presence of entities in the input text also plays a crucial role in extraction. As the distance between entities increases, the complexity and vagueness also increase, which makes the prediction harder. *Dsip and Clip [acth(18-39)] immunoreactive neurons and fibers were examined in the **human** hypophysis and pituitary stalk using immunohistofluorescence and **peroxidase** antiperoxidase methods.* Here, the actual relation between human and peroxidase is **is_organism_source_of_gene_product**.

To identify the relationship between two entities in this sentence requires more than just an input sentence. Our work delves into the dynamic and evolving field of knowledge-enabled relation extraction (KERE), a multifaceted approach designed to harness the rich contextual cues of entities from public knowledge bases to predict the relations between two given entities [15]. Unlike conventional RE models, KERE systems operate on diverse text formats, ranging from individual sentences to entire documents and even across multiple documents. This adaptability allows KERE to cater to a wide array of real-world scenarios where entities of interest may be discussed in varying levels of detail, depth, and context.

Very few studies in the literature investigate how background knowledge affects the RE task. Background knowledge can provide a deeper context that helps in accurately interpreting the meaning of entities and their relationships. For example, understanding that "Mumbai" can refer to a city in India or a person’s name depending on the context can drastically change the extracted relationship. Background knowledge can also help disambiguate terms and ensure that the entities and relationships are correctly identified. For instance, knowing that "Jaguar" can refer to an animal, a car brand, or a software name or a football team is essential for accurate relation extraction in different domains. Also, different domains have unique terminologies,

concepts, and common patterns of relationships. Background knowledge about a specific domain (e.g., medicine, market, finance) can guide the relation extraction process to be more aligned with domain-specific expectations, improving the relevance and accuracy of the results. With background knowledge, systems can infer relationships that are not explicitly stated but are implied by the context. This ability to read between the lines and understand implied relationships is crucial for comprehensive relation extraction. Additionally, background knowledge enriches the relation extraction process by providing a deeper, better understanding of the text. It enables more accurate, relevant, and comprehensive extraction of relationships, which is crucial for a wide range of applications in information retrieval, knowledge management, decision support systems, and beyond.

In this dissertation , we study the significance of background knowledge within relation extraction systems [16]. Not only do we showcase the advantages of integrating background knowledge to enhance system performance, but we also introduce a structured framework designed to effectively integrate knowledge into statistical deep learning models for relation extraction across various types of input text. We discuss a novel framework and a model that integrates ontology as an external context to extract relations from a sentence. Following this, we discuss another framework capable of utilizing documents rather than individual sentences to predict relationships. Thirdly, we provide an overview for identifying relationships across multiple documents. In each scenario, we present evidence showcasing the enhancement in the performance of deep learning methods by integrating the external knowledge of entities.

1.3 Existing Gaps

The existing approaches employ various techniques such as Multi-task learning [17], Transformers [18], and Graph Neural Network (GNN) models [19, 20] to process complex relationships between entities. Deep learning models [21, 22] can incorporate semantic information of entities. Albeit effective, these models employ standard message-passing or attention-based approaches (Transformers, GNNs) which are inherently focused on homophilic signals [23, 24] (i.e., only on neighbourhood interactions) and ignore long-range interactions that may be required to infer the semantic relationship between two entities. Furthermore, sufficient domain-specific knowledge is available in various knowledge bases such as Wikidata [25], DBpedia [26], and ontology repositories [27] to be used as background knowledge for relation extraction. It is also evident in the literature that reasoning over knowledge bases [28, 29] allow capturing long-range dependencies between two entities [30, 31], which further helps in making predictions. For instance, in [32], ontology information was utilized as a tuple and transformed into a 3-D vector for predicting compound relations.

1.4 Problem Formulation and Statement

We define a Graph as a tuple $G = (\mathcal{E}, \mathcal{R}, \mathcal{T}^+)$ where $\mathcal{E}$ denotes the set of entities (vertices), $\mathcal{R}$ is the set of relations (edges labels), and $\mathcal{T}^+ \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is a set of all triples. The *RE Task* aims to find the target relation $r^c \in \mathcal{R}$ for a given pair of entities $\langle e_i, e_j \rangle$ within the sentence $\mathcal{W}$, document $\mathcal{D}$ and across multiple documents $\mathcal{C}_{i=1}^{n}$(n is total number of documents). If no relation is inferred, it returns the *NA* label.

Our objective is to uncover the intricate connections between entities, ranging from overtly stated facts to implied correlations from input text. The input text can be a sentence, a document, or multiple documents. To achieve this objective, we study the impact of external knowledge on the relation extraction task. We provide a deep learning-based framework employing external knowledge, which enhances the performance of the relation extraction task.

1.5 Research Objectives

The primary aim of this dissertation is to identify relationships within the text, by considering the granularity of the input text. Our hypothesis is that incorporating structured domain knowledge about entity types and their relationships will enhance the performance of relation extraction models. We argue that relation extraction models can be improved with one or multiple sources of background knowledge and can significantly enrich the depth of model inputs and the bias of final outputs. Our objectives are divided into three distinct parts based on the granularity of the input text.

- **Sentence-level relation extraction(RE)** is the task of extracting semantic relationships between entities in a sentence. We study the impact of external knowledge in the form of expressive axioms present in ontology in these RE tasks, which is discussed in chapter 4.
- **Document-level relation extraction (DocRE)** poses the challenge of identifying relationships between entities within a document as opposed to the traditional RE setting where a single sentence is an input. We observe the performance of DocRE tasks with external knowledge (Knowledge Graph), which is discussed in chapter 5.
- **Cross document-level relation extraction (CrossDocRE)** is a relation extraction task between entities that can extend across multiple documents. Compared to RE and DocRE, it is more complex and challenging. We extend our study in this setting to identify the usefulness of integrating domain knowledge. Details are discussed in chapter 6.

1.6 Thesis Related Publications

- Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part IV. Springer-Verlag, Berlin, Heidelberg, 230–247. https://doi.org/10.1007/978-3-031-43421-1_14
- Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024, Proceedings of the AAAI Conference on Artificial Intelligence. 38, 16 (Mar. 2024), 18327-18335. https://doi.org/10.1609/aaai.v38i16.29792.
- Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024. 3787–3797. Findings of the Association for Computational Linguistics ACL 2024 10.18653/v1/2024.findings-acl.227
- Monika Jain. 2024. Knowledge Enabled Relation Extraction. PhD Symposium Track. In Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3589335.3651263

1.7 Other Publications

1. Monika Jain, Paramita Mirza, and Raghava Mutharaju. 2020. Cardinality Extraction from Text for Ontology Learning. Young Research Symposium Track. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020). Association for Computing Machinery, New York, NY, USA, 354. https://doi.org/10.1145/3371158.3371223
2. Nikhil Sachdeva, Monika Jain, and Raghava Mutharaju. 2021. Extraction of Union and Intersection Axioms from Biomedical Text. Demo Track. In The Semantic Web: ESWC 2021 Satellite Events: Virtual Event, June 6–10, 2021, Revised Selected Papers. Springer-Verlag, Berlin, Heidelberg, 147–151. https://doi.org/10.1007/978-3-030-80418-3_27

1.8 Organization

The rest of the dissertation is organized as follows: In Chapter 2, we review existing literature for relation extraction along with their challenges and applications. In chapter 3, we discuss Symbolic AI, Subsymbolic/Connectionist AI, and NeuroSymbolic AI. In Chapter 4, we discuss the extraction of relation between entities present in a sentence. In Chapter 5, we focus on

identifying the relation between entities present in a document containing multiple sentences. In Chapter 6, we discuss the problem of relation extraction between entities across documents. We conclude our dissertation in Chapter 7.

CHAPTER 2

Related Work

Relation extraction is a fundamental task in natural language processing (NLP) that focuses on identifying and categorizing the semantic relationships that exist between two entities mentioned in the text. This task has gained significant attention due to its crucial role in various NLP applications, such as information retrieval, knowledge graph completion, question answering systems and ontology learning. Over the years, extensive research has been conducted to develop and improve relation extraction techniques.

2.1 Introduction

Recent statistics [33] reveal that in the year 2020, global data consumption totalled 64.2 zettabytes. Projections anticipate this figure to exceed 180 zettabytes by 2025. To fully harness this data, it requires conversion from unstructured to structured form, enabling utilization in question answering, knowledge graph completion, and information retrieval [34, 35, 36]. The primary steps in this conversion process entail named entity recognition (NER) and relation extraction (RE). NER is a sub task in information extraction that seeks to locate and aclassify named entities mentioned in unstructured text into categories such as person name, location and quantities among others. After NER identification, RE predicts the semantic relationship between two named entities. Although entity recognition methods have been proposed and tested with satisfactory accuracy [37], relation extraction still remains a challenge. There is a need for a generalized method applicable across all the domains to address this hurdle. Due to its vast amount of usage in various applications, RE task has been popular for decades and still it is of keen interest to researchers. The task of relation extraction can be addressed in two ways: either by directly predicting the relation between two given entities [38], or through a pipeline approach where entities are first identified and then the relation extraction is performed between them [39]. Both methods are widely utilized and chosen based on the specific application's requirements. We categorize the relation extraction task based on the granularity of the text, which can be a sentence [40, 41, 42, 43, 44, 15], document [45, 46, 47, 48, 49, 50, 51] or across documents [52]. When predicting the relationship between two entities within the text, it is classified as follows: if the text is a sentence, it's referred to as sentence-level RE; if it's a document, it's termed document-level RE (DocRE); and if there are multiple documents, it's known as cross-document RE (CrossDocRE).

An example of sentence-level RE is shown in Figure 2.1. Given the sentence and two entities *GE Aerospace* and *NASA*, the relation label between the two entities is *collaborates with*. Similarly, the relation label between *GE Aerospace* and *aerospace technologies* is *works on*. Likewise, there are other relation labels between entities that can be extracted from the text. For document-level RE, Table 2.2, shows the document consisting of sentence from 1 to 11 and Figure 2.3, shows the relation label between two entities and evidence sentence number from which the relation is inferred in a document. Let us consider the first example where the entities are *Royal Court Orchestra* and *Royal Swedish Opera*, and the task is to predict the relation label as *part of*. The evidence shows the sentence number referred to make the prediction. Likewise, for entity *Stockholm* and *Sweden*, the task is to predict the relation label as *capital of*. For cross-document relation extraction, consider Figure 2.2. Multiple documents are provided here. We first search for the source entity in the documents. Likewise, we search the documents for the target entity. We then retrieve all relevant paths via the bridge entities (entities other than the source and the target entity) connecting these entity pairs. In this example, *GCompris* and *GNU Project* are the source and target entities connected via different reasoning paths containing bridge entities such as *GNU, Linux and Qt*.

2.2 Related work

Since sentence-level relation extraction is a popular approach, there are a few survey papers that discuss RE tasks and methods. One of the first surveys discusses RE methods based on categories of supervised, semi supervised and unsupervised methods from the year 1995 to 2007 [53]. While it includes fundamental methods and techniques, it lacks the latest advancements in this domain. Later, other surveys such as [54] cover detailed descriptions of supervised, unsupervised and distant supervised methods for RE tasks from the year 1998 to 2014. Another survey includes the state-of-the-art RE work done in the Portuguese language [55]. It discusses the different aspects related to the task, considering the main computational strategies, used resources, as well as the evaluation methods applied. Additionally, survey papers that discuss RE methods focusing on deep learning methods [56, 57, 58] and distant supervision methods [59] in detail. This is the first dissertation which discusses RE tasks based on the granularity of the text.

Biomedical sentence-level RE: Biomedical relation extraction has been a popular task for the past decade, with several comprehensive surveys conducted on researching this field. The first survey paper in this domain discusses a general framework for biomedical relation extraction from the year 1989 to 2014 and also proposes an approach for binary and complex relation extraction [60]. In another survey paper, an overview of biomedical relation extraction using

distant supervision is discussed in detail from the year 2003 to 2020 [61].

There are a few survey papers which discuss document-level RE in detail. In one of the methods, a comprehensive survey of document-level relation extraction from the year 2016 to 2023 was discussed [62]. In another survey paper, different relation extraction models on document-level RE, along with different datasets and evaluation metrics, are discussed [63]. As of now, there haven't been any survey papers discussing cross-document-level relation extraction. This is the first dissertation that discusses a survey on cross-document relation extraction. Table 2.1 provides the details of the surveys done so far. The table includes details such as the granularity level of text, source, year in which it was published, topics covered in the survey and years covered.

Table 2.1: Previous survey papers with description

Granularity level	Year	Source	Years covered	Topics covered
Sentence	2007	[53]	1995-2007	supervised, semi-supervised method
Sentence	2007	[54]	1998-2014	supervised, semi-supervised, unsupervised, distant-supervision
Sentence	2013	[55]	1992-2012	relation extraction on Portuguese, developed systems for Portuguese
Sentence	2017	[56]	2007-2017	deep learning methods
Sentence	2022	[57]	1992-2021	deep learning methods
Sentence	2021	[58]	1987-2019	NER and overview of relation extraction methods
Sentence	2019	[59]	2002-2018	distant supervision method
Biomedical sentence	2014	[60]	1989-2014	rule based, machine learning method
Biomedical sentence	2020	[61]	2003-2020	distant supervision method
Document	2023	[62]	2016-2023	Graph and transformer method
Document	2023	[63]	2016-2022	relation extraction techniques, datasets, evaluation

Contribution of this chapter:

1. We provide a detailed survey of the relation extraction techniques, covering their evaluation methodologies and datasets.
2. We provide insights of relation extraction with respect to the text granularity level, i.e., sentence-level, document-level and cross-document-level.
3. For each dataset with respect to the granularity level, we analyse the most promising model proposed in the existing literature.

This chapter is organized as follows. In section 1, we detail the examples of sentence-level, document-level and cross-document-level RE. In the section 2 and 3, we review the existing survey on relation extraction tasks. In the section 4, we briefly understand the RE based on rule based, supervised, unsupervised, distant supervision, unsupervised and knowledge based methods. So far, this is the first work that discusses knowledge based methods in detail. In sections 5 and 6 , we discuss datasets available on RE task based on sentence-level, document-level and cross-document-level, along with the evaluation metrics. The evaluation results are discussed in the seventh section. We also discuss the best model available for each dataset. In

Section 8, we discuss the use of large language models for RE. In Section 9 and 10, we discuss the applications and challenges of the RE task.

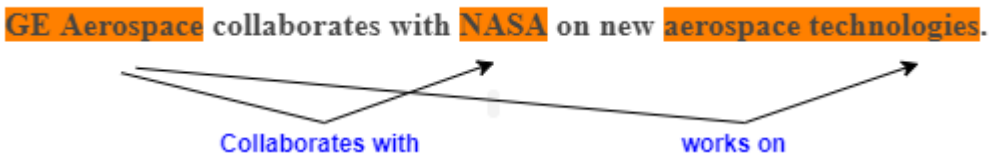

Figure 2.1: An example of sentence-level relation extraction.

No.	Sentence
1	Kungliga Hovkapellet (The Royal Court Orchestra) is a Swedish orchestra originally part of the Royal Court in Sweden's capital Stockholm
2	Its existence was first recorded in 1526
3	Since 1773 it is part of the Royal Swedish Opera's company.
4	Kungliga Hovkapellet is one of the oldest active orchestras in the world
5	It was first recorded in the royal account books from 1526
6	The orchestra originally consisted of both musicians and singers
7	It had only male members until 1727 when Sophia Schröder and Judith Fischer were employed as vocalists; in the 1850s the harpist Marie Pauline became the first female instrumentalist
8	It had a golden age under the leadership of several members of the Düben family during the 17th century
9	In the 18th century its directors included Johan Helmich Roman and Francesco Uttini
10	From 1731 public concerts were performed at Riddarhuset in Stockholm
11	Since 1773 when the Royal Swedish Opera was founded by Gustav III of Sweden the Kungliga Hovkapellet has been part of the opera's company.

Table 2.2: Sentence Information

Entity 1	Entity 2	Relation	Evidence
Royal Court Orchestra	Royal Court	part of	1,2,10
Royal Swedish Opera	Royal Court Orchestra	subsidiary of	1,10,2
Marie Pauline	Royal Court Orchestra	employer	1,6
Sweden	Stockholm	capital	1
Stockholm	Sweden	capital of	1

Table 2.3: Entity1 and Entity2 represents the source and target entity and Relation represents the relation between two entities. The evidence shows the sentence number referred to make the prediction

2.3 History of Relation extraction

Relation extraction originated in 1980s from the Message Understanding Conferences (MUC). MUC conference is initiated by the Defence Advances Research Projects Agency (DARPA) to encourage the development of novel information extraction methods. Details of the conferences are given in Table 2.4. In each MUC, new techniques developed for different IR problems are

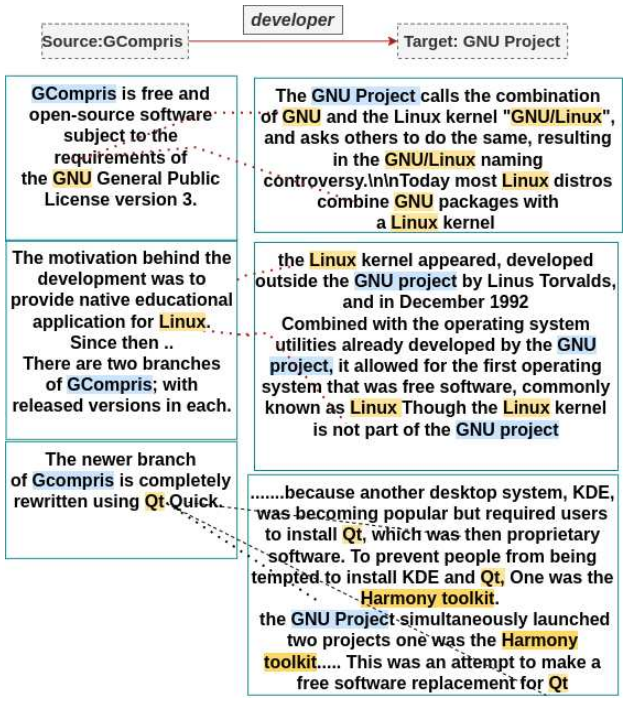

Figure 2.2: An example of cross-document-level relation extraction. Blue colour represents named entities, and yellow colour represents mentioned entities

discussed. In 1998, the relation extraction task is discussed, which involves the identification of relations named: location_of, employee_of, and product_of. Additionally, MUC-7 also proposed evaluation techniques such as Precision, Recall and F1-score. MUC conferences have significantly contributed to the advancement of relationship extraction research. The National Institute of Standards and Technology (NIST) launched the automatic content extraction (ACE), succeeding MUC. This program was held from year 1999 to 2000. ACE program also addressed the same issues as MUC, however it defined the research objectives in terms of target objects such as the entities, the relations and events. With the MUC and ACE events becoming popular, relational extraction technology has seen remarkable progress. Research trends have evolved from applying simple language models to utilizing NLP technology with shallow parsers or full parsers, and further to employing complex machine learning methods. As a result, the performance of relation extraction has greatly improved.

2.4 Methods

We discuss the RE methods based on the text granularity as shown in Figure 2.3.

Table 2.4: Details of previous conferences depicting the history of relation extraction. Here IE, CR, NER represents information extraction, co reference resolution, named entity entity recognition

Conference	Year	Task	Source text
MUC-1	1987	military operations	military reports
MUC-2	1989	military operations	military reports
MUC-3	1991	IE from template	news reports
MUC-4	1992	IE from template	news reports
MUC-5	1993	IE from template	news reports
MUC-6	1998	NER, CR and scenarios	news reports
MUC-7	1998	CR, NER, event extraction, scenarios	news reports
ACE-Pilot	2000	NER	news reports
ACE-1	2001	NER	news reports
ACE-2	2002	NER, RE	news reports
ACE 2003	2003	NER, RE	news reports
ACE 2004	2004	NER, RE, events	news reports

2.4.1 Sentence-level RE

Sentence-level relation extraction aims at identifying the relationship between two entities in a sentence [64]. Several methods have been proposed for sentence-level relation extraction, including rule based, supervised, unsupervised, distant supervision, and knowledge-based approaches.

Problem Definition: We define a Graph as a tuple $G = (\mathcal{E}, \mathcal{R}, \mathcal{T}^+)$ where $\mathcal{E}$ denotes the set of entities (vertices), $\mathcal{R}$ is the set of relations (edges labels), and $\mathcal{T}^+ \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is a set of all triples [19, 38]. The *RE Task* aims to find the target relation $r^c \in \mathcal{R}$ for a given pair of entities $\langle e_i, e_j \rangle$ within the sentence $\mathcal{W}$. If no relation is inferred, it returns *NA* label.

In this section, we will explore each method related to sentence-level RE in detail.

Rule Based Methods

Rule-based methods are a type of algorithm that utilize pre-defined rules to classify data or make predictions, instead of learning rules from data.

In a prior study, a rule-based relation extraction system was developed using DBpedia and syntactic parsing [65]. To define binary relationships, the researchers employed two distinct types

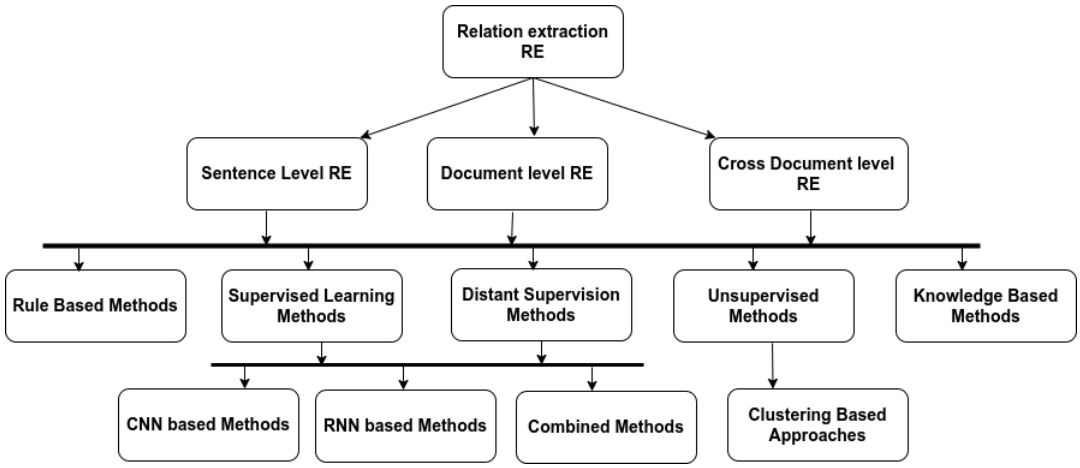

Figure 2.3: Illustration of categories of relation extraction methods

of relation extraction patterns: lexical patterns, constructed from words and word class information (e.g., "X was born in Y"), and dependency patterns that incorporate syntactic information (e.g., "subject-verb-object" structures). Let's consider an example where we want to predict the presence of the "works for" relation using a rule based approach. We can use the keyword matching method to identify someone who is working for an organization. For this specific example, keywords can be "works for", "employed by", "job at", etc. Sentence pattern can be *PERSON [verb] for ORGANIZATION* where PERSON is a proper noun tagged as a person's name, and ORGANIZATION is tagged as an organization name. Dependency parsing can be used to analyse the dependency relationship between words in the sentence. For example, identify the pattern like "PERSON [verb] for ORGANIZATION" where PERSON is the subject of the verb and ORGANIZATION is the object.

For the sentence: "Rebecaa works for xyz Corporation". Rule 1: Using keyword matching "works for" keyword is identified in the sentence. Rule 2: POS tagging identifies "Rebecaa" as PERSON and "xyz Corporation" as ORGANIZATION. Rule 3: Dependency parsing identifies "Rebecaa" as the subject, "works" as the verb, and "xyz Corporation" as the object. Using these rules, the relation extraction system can confidently extract the relation "works for" from this sentence.

Supervised Learning Methods

Supervised methods need a large amount of labelled data to train a classifier to predict the relation. There are two types of supervised learning approaches, `feature based methods` and `kernel based methods`. In feature based approaches, feature data is generated for each relation, and a classifier is used for training a relation instance [54, 66, 67, 68]. One of the feature based methods employed a maximum entropy model to combine diverse lexical,

syntactic and semantic features derived from the text to obtain the content from text [67]. In addition, an extension of this work incorporates WordNet [69] and name list, to further improve the performance [68]. It is important to select features carefully as they directly impact the performance. While good features can improve the performance, bad features can degrade the performance. To solve this problem, kernel based methods are introduced to exploit the rich representation of the input data [70].

In kernel based methods, kernel functions are designed to compute similarities between representations of two relation instances and SVM (Support Vector Machines) is employed for classification [71, 72, 70, 39, 73]. For the supervised approach, labelled dataset was used where each sentence was annotated with a relation between entities. String kernels have been introduced and discussed in the context of text classification [74]. Given two strings x and y, the string-kernel computes their similarity based on the number of sub sequences that are common to both of them. The more the number of sub sequences common, the greater the similarity between the two strings. Each string can be mapped to a higher dimensional space where each dimension corresponds to the presence (weighted) or absence of a particular sub sequence.

Let us understand the supervised method (feature based) using the previous example, "Rebecaa works for xyz Corporation" and the annotated sentence is "Rebecaa, works for, xyz Corporation". For each sentence, relevant features are extracted that can help the model learn the relation. These features include: a) lexical features, which include words in the sentence, their part-of-speech tags, dependency parse tree features etc. b) contextual features, which include surrounding words or phrases, sentence structure etc. c) entity features, which include types of entities involved, their positions in the sentence etc. A supervised machine learning model is trained using these features. The features of the example sentence can be,

1. Lexical features: ["Rebecca", "works", "for", "xyz", "Corporation"]
2. POS tags: ["NNP", "VBZ", "IN", "NNP", "NNP"]
3. Dependency parse tree features: [subject: "Rebecca", verb: "works", object: "xyz Corporation"]
4. Contextual features: ["Rebecca" and "xyz Corporation" appearing in close proximity, indicating a potential relation]
5. Kernel features: A kernel function which computes similarity between pairs of feature vectors.

Based on these features, the trained model predicts the relation "works for" between "Rebecca" and "xyz Corporation"

Unsupervised Methods

Unsupervised methods do not use any labelled data. Clustering algorithms is an example of unsupervised learning [75, 76]. To identify potential relationships between entities in the text using an unsupervised approach, basic techniques such as entity recognition, context analysis, co-occurrence statistics, association measures, thresholding, and clustering are generally used, without relying on manually annotated data. Consider the same example “Rebecca works for XYZ Corporation. XYZ Corporation is headquartered in China. Rebecaa frequently visits the China office”.

1. (Rebecca, XYZ Corporation): Identified due to their frequent appearance together in the text.
2. (XYZ Corporation, China): Identified due to their frequent appearance together in the text.
3. (Rebecca, China): Identified based on the frequent co-occurrence of "Rebecca" and "China" in discussions related to XYZ Corporation’s headquarters.

In this example, the unsupervised approach identifies potential relations between entities based on their co-occurrence patterns in the text without relying on manually labelled training data. This means that the method looks for instances where different entities (such as people, places, or things) appear together frequently within the same context. By examining these patterns, the approach can infer that there might be a relationship between these entities. However, assigning specific names to these relations requires additional information or a separate process.

Distant Supervision Methods

The distant supervision method is a popular approach that generates training data automatically using a knowledge base, addressing the challenge of labeling training data [77]. The knowledge base employed in distant supervision helps in identifying the relationship between two entities. If both a sentence and a knowledge base contain the same pair of entities, this implies that the relation between them is heuristically linked with the sentence.

Consider an example. We start with a large knowledge base such as Freebase, DBpedia, or Wikidata, which contains structured information about entities and their relations. For example, Freebase might contain information that “Rebecca” works for “XYZ Corporation”. If two entities appear in the same sentence and are related in the knowledge base, we label that sentence with the corresponding relation. For example, if “Rebecca” and “XYZ Corporation” appear in a sentence together, and they have a “works for” relation in the knowledge base, we label that sentence as “Rebecca works for XYZ Corporation”.

Following this initial labeling, features (lexical, syntactic, semantic) are extracted from the

labeled data. These features are then used to train a supervised model. The model learns to generalize from the heuristically labeled data, improving its ability to predict relations in new, unseen sentences. This process iteratively refines the model's performance, making distant supervision a powerful technique for relation extraction tasks.

Knowledge Based Methods

Knowledge based methods integrate background knowledge with the text and feed it to the model [78, 79]. Let us take the same example as earlier, "Rebecca works for xyz Corporation". Let us consider the entity type as an external knowledge. The entity type of "Rebecca" is PERSON, and the entity type of "xyz Corporation" is ORGANIZATION. When this information is fed into the model, considering both entity types, the relation "works for" is more viable than other relations. Other than entity type, there are various types of background knowledge, such as ranking relations, which can be fed into the model to improve its performance.

Existing knowledge based approaches use entity types and entity aliases to predict the relation [78, 79]. RECON [80] encoded attribute and relation triples in the Knowledge Graph and combined the embedding with their corresponding sentence embedding. In contrast to these methods, we proposed the ReOnto model that considers external knowledge as a context path between entities retrieved using ontology to predict the relation [15].

2.4.2 Document-level RE

Previous efforts [81, 44] in relation extraction have focused on predicting relationships within a single sentence. Recent years have seen growing interest in relation extraction beyond a single sentence, called document-level relation extraction (DocRE) [82, 83, 84, 81, 12]. In contrast to sentence-level relation extraction, document-level relation extraction (DocRE) poses greater complexity and challenges. Significant efforts have been dedicated to the task of extracting relations at the document-level as well.

Problem definition:

An unstructured document D consisting of K sentences is represented by $\{S\}_{i=1}^{k}$, where each sentence is a sequence of words and entities $\mathcal{E}$=$\{e_i\}_{i=1}^{P}$ (P is the total number of entities). Entity e_i has multiple mentions, $m_i^{s_k}$, scattered across the document D. A mention $m_i^{s_k}$ refers to a specific occurrence or reference to the entity e_i within the sentence s_k. These mentions are the instances where the entity is explicitly mentioned in the text, and they help in identifying and linking the entity throughout the document. An entity alias is represented as $\{m_i^{s_k}\}_{k=1}^{Q}$ where Q represents the total number of mentions for a specific entity e_i within the document

D. Our objective is to extract the relation between two entities in $\mathcal{E}$ namely P(r|e$_i$,e$_j$) where $e_i, e_j \in \mathcal{E}, r \in \mathcal{R}$, here $\mathcal{R}$ is a total labelled relation set. The context (background knowledge) of an entity e_i is represented by C_{e_i} and a context path, i.e., a sequence of connected entities and edges from the head entity (e$_i$) to the tail entity (e$_j$) is represented by CP_{e_i,e_j}. In this section we will explore each method in detail.

Rule Based Methods

A probabilistic model (LogiRE) for DocRE was introduced, which learns by logic rules [85]. LogiRE treats logic rules as latent variables consisting of two modules: a rule generator and a relation extractor. The rule generator creates logic rules that may influence the final predictions. The relation extractor then uses these generated logic rules to produce the final predictions.

In another study, heuristic rules are designed to select informative path sets from documents, which can be further combined with a simple BiLSTM to achieve performance on a benchmark dataset [86]. Similarly, previous studies have utilized heuristic rules [47, 49], or syntactic dependency paths [87, 88] to construct a pseudo graph structure for their GNN models, establishing binary associations between pairs of entities based on task-independent auxiliary information.

Supervised Learning Methods

The majority of literature on document-level relation extraction (DocRE) relies on a supervised approach for extraction. This approach encompasses sequence-based, graph-based, and transformer-based methodologies. Here are the details of these approaches.

Sequence Based Approach:

Sequence-based approaches view the document as an augmented sequence and apply a sequential model derived from sentence-level relation extraction to identify relations between specific entities. For extracting Chemical-Induced Disease (CID) relations on the CDR dataset, a maximum entropy system is used for inter-sentence relations, while a CNN model is applied for intra-sentence relations [89]. Recurrent piecewise convolutional neural networks (RPCNN) are used for CID extraction, combining domain knowledge, piecewise strategy, attention mechanism, and multi-instance learning [26]. In another study, combining RNN and CNN was proposed. This exploits word embeddings and positional embeddings for cross-sentence n-ary relation extraction [90]. In another method, RENET was introduced, which combines a CNN and an RNN with gated recurrent unit (GRU) gates for gene-disease association extraction [91].

Graph Based Approach: The graph approach is based on a graph constructed with mentions, entities, sentences, or documents. The relations between these nodes are then deduced through

reasoning on this constructed graph. Earlier work uses multi-hop reasoning on a mention level graph for inter-sentential entity pair [49]. In another work, a proposed model named Grail reasons over local sub-graph structures and has a strong inductive bias to learn entity-independent relational semantics [92]. A discriminative reasoning framework (DRN) was introduced in a different study. This framework involves modeling the pathways of reasoning skills connecting various pairs of entities [93]. DRN was designed to estimate the relation probability distribution of different reasoning paths based on the constructed graph and vectorized document contexts for each entity pair, thereby recognizing their relation.

Transformer Based Approach: It is another interesting approach to tackle the document-level relation extraction problem [49]. The primary focus revolves around maximizing the effective utilization of long-distance token dependencies using a transformer. One of the approaches considers DocRE as a semantic segmentation task, employing the entity matrix, and they utilize a U-Net to capture and model [94]. In a separate study, localized contextual pooling was introduced to focus on tokens relevant to individual entity pairs [95]. Another approach incorporates explicit supervision for token dependencies, achieved by leveraging evidential information [96].

Distant Supervision Methods

Distant supervision was a popular method in traditional RE tasks. An approach [97] uses a novel pre-trained model to denoise the document-level distant supervision data via multiple pre-training tasks. In another study, to augment document-level relation datasets, a method integrating an LLM and a natural language inference (NLI) module has been proposed [98].

Knowledge Based Methods

Recently, there has been a growing trend towards integrating external knowledge into Document-Level Relation Extraction (DocRE) tasks, which has shown promising results. KB-Both [99] uses entity details from hyperlinked text documents from Wikipedia and Knowledge Graph (KG) from Wikidata to enhance performance. Another approach [100] integrates knowledge, including co-references, attributes, and relations with different injection methods, for improving the state-of-the-art.

2.4.3 Cross-document-level RE

The relation between entities can extend across multiple documents, and researchers have investigated the extraction of entities, events, and relationships in a cross-document context with unlabelled data [101]. Cross-document relation extraction (CrossDocRE) has been gaining attention. The first human annotated dataset has been developed, and it is called CodRED [52].

To date, the majority of CrossDocRE related research has employed a supervised approach. Notably, there is a lack of exploration in this area using alternative methods such as rule based and unsupervised approaches.

Supervised Approach

CrossDocRE problem has been approached using two methods. The first method involves a pipeline approach in which a relational graph is constructed for each document and subsequently reason over these graphs to extract the desired relation. The second method, referred to as the joint approach, directly combines various path representations through a selective attention mechanism to predict relations. An effective cross-document relation extraction (DocRE) system should incorporate bridge entities to establish connections between entities. However, in their work [52], the authors only superficially consider text paths and do not take into account bridge entities or multi-hop reasoning for relation prediction. Considering these challenges, an entity-centered cross-document relation extraction (ECRIM) technique was proposed. It uses an entity-based document context filter to retain useful information in the given documents by using the bridge entities in the text paths. Secondly, they solve CrossDocRE using cross-path entity relation attention, which allows the entity relations across text paths to interact with each other [102]. Nevertheless, this work focuses on the closed setting where evidential context has been given instead of all documents. Following this work, multi-hop evidence retrieval for cross-document relation extraction (MRCOD) was proposed, which implements a multi-hop evidence retrieval method based on evidence path mining and ranking. A multi-document passage graph, where passages are linked by edges corresponding to shared entities, is constructed. Then, a graph traversal algorithm is used to mine passage paths from head to tail entities. Paths are ranked based on relevance, and top-K evidence paths are selected as input for downstream relation extraction models. A causality guided algorithm is used to filter confusing information and achieve global reasoning to solve cross-document relation extraction [103].

Distant supervision method

Distant supervision is used to train a factor graph model for relation extraction based on an existing knowledge base (Freebase, derived in parts from Wikipedia) [104]. This is the first work that extracts the relation in cross-document setting using a distant supervision method. For inference, an efficient Gibbs sampler with linear time joint inference is used. This approach is evaluated on an in-domain (Wikipedia) and a more realistic out-of-domain (New York Times Corpus) setting.

2.5 Datasets

Table 2.5 shows the commonly used datasets for sentence-level RE. In RE tasks, the dataset plays a crucial role in developing and testing the model. A large scale annotated corpus is required to train and evaluate the RE systems. The datasets are generated either manually, or by crowd sourcing or derived from other datasets. Most of the machine learning and deep learning based models rely on size and quality of datasets. Due to this reason, other than manual annotation, NLP based techniques are also used for creating or extending the datasets. Following are the details of popular datasets used in this field.

CoNLL04: This dataset was released in the year 2004. There are four entity types in this dataset (Organisation, Location, Person and Other) and five relation types (Live in, Kill, Work for, Located in, OrgBased in). This dataset consists of 910 samples in the training set, 243 in the validation set and 288 for the testing set [105]. This dataset is constructed using news articles.

ACE2004: The Automatic Content Extraction 2004 (ACE 2004) is a multilingual training corpus containing the complete set of English, Arabic and Chinese training data [106]. There are seven entity types, namely Person, Organization, geographical entities, Location, Weapon, Facility, and Vehicle. This dataset contains 7 relation labels, namely Physical, Person-Social, Agent-Artifact, GPE affiliation, Employment-Membership-Subsidiary, Other affiliation, and Discourse.

ACE2005: The Automatic Content Extraction 2005 (ACE2005) dataset contains annotated events and relations in English, Arabic, and Chinese languages [107]. It contains 599 English documents from various sources, such as news, forums, and telephone conversations. It contains 7 major relation types, with an average of approximately 700 examples per relation.

NYT: The New York Times (NYT) dataset [108] is a large collection of articles published between 1996 and 2007. This dataset contains 24 types of relation labels that may occur in a sentence. Each sentence can have more than one relation.

SemEval-2010 Task-8: The SemEval-2010 Task 8 focuses on the multi-way classification of semantic relations between pairs of nominals. This dataset was released in the year 2010. The task was designed to compare different approaches to semantic relation classification and to provide a standard test bed for future research.

ADE: The adverse drug events (ADE) corpus (released in the year 2012) is a benchmark corpus to support the automatic extraction of drug related adverse effects from medical case reports. There are 6821 sentences in this dataset. It is a gold standard corpus created from MEDLINE case reports.

Table 2.5: Overview of the datasets used for sentence-level RE

Datasets	Paper	Year	Evaluation metrics	Relation types	Sentences/ Examples	Source
CoNLL04	[105]	2004	F1	5	1441	News articles
ACE2004	[106]	2004	F1	7	-	news articles
ACE2005	[107]	2005	F1	6	5349	News and conversations
NYT	[108]	2008	P, R, F1	24	61195	New York Times Articles
SemEval-2010 Task-8	[109]	2010	Macro F1 score	9	10717	General
ADE	[110]	2012	F1	2	6821	Biomedical text

Document-level RE datasets are created to identify the relations of various entity pairs expressed across multiple sentences. Compared to sentence-level RE, document-level RE datasets are more complex and challenging. Table 2.6 presents a summary of key statistics for each dataset, including the year of publication, evaluation metrics used, number of relation types, total number of documents and source of the dataset. We provide details of some popular DocRE datasets.

ChemProt: ChemProt consists of 1,820 PubMed abstracts with chemical-protein interactions annotated by domain experts and was used in the BioCreative VI text mining chemical-protein interactions shared task [111]. Chemprot is released in the year 2011 and contains 14 relation types.

CDR: The BioCreative V CDR task corpus (released in the year 2016) is manually annotated for chemicals, diseases and chemical-induced disease (CID) relations. It contains the titles and abstracts of 1500 PubMed articles and is split into equally sized train, validation and test sets [112].

DocRED: DocRED (Document-level relation extraction dataset) is constructed from Wikipedia and Wikidata [83]. Each document in the dataset is human-annotated with named entity mentions, co-reference information, intra and inter-sentence relations, and supporting evidence. DocRED contains 132,375 entities and 56,354 relational facts annotated on 5,053 Wikipedia documents. DocRED is one of the most popular datasets to date for DocRE task. It is released in the year 2019.

GDA: GDA is a gene disease associations corpus containing 30,192 titles and abstracts from

PubMed articles that have been automatically labelled for genes, diseases and gene-disease associations via distant supervision. The test set consists of 1,000 examples, and the validation set contains 20% of the training set data [113].

DWIE: Deutsche Welle corpus for Information Extraction (DWIE) is a multi-task dataset that combines four main Information Extraction (IE) annotation sub-tasks: (i) Named Entity Recognition (NER), (ii) Coreference Resolution, (iii) Relation Extraction (RE), and (iv) Entity Linking. DWIE is constructed from Deutsche Welle articles and released in the year 2019 [114].

HacRED: HacRED is a large scale Chinese relation extraction dataset for practical applications [115]. HacRED released to include practical hard cases to make DocRE tasks more challenging. HacRED consists of 65,225 relational facts annotated from 9231 wiki documents with sufficient and diverse hard cases.

ReDocRED: ReDocRED is the revised version of the dataset released to solve the problems of DocRED [116]. It addresses the incompleteness, inconsistencies and coreferential errors of DocRED. The ReDocRED dataset is released in the year 2022.

CodRED: CodRED is the first and the only dataset available for cross-document-level relation extraction [52]. After its release, CrossDocRE started to be a new emerging subtask of RE. The statistics of CodRED are given in Table 2.7. CodRED is constructed using Wikipedia and its official evaluation metric is F1 score. It contains 276 relation types. It was released in the year 2021. It contains an average of 3646 reasoning paths for each entity pair.

2.6 Evaluation metrics

Precision, recall and F1 score [118] are the metrics generally used to measure the performance of the RE systems. Precision measures the proportion of correctly classified positive samples among all positive predictions made by the model. Recall is a metric which measures how often a machine learning model correctly identifies positive instances from all the actual positive samples in the dataset. The F1 score represents the harmonic mean of the precision and recall of a classification model. Micro average F1 score involves calculating the total true positives, false positives, and false negatives across all classes and then computing precision, recall and F1 score. Macro average calculates the average of F1 scores for each class without considering class imbalance. Ign F1 is an F1 score excluding relational facts that exist in both training and development/testing sets. Ign F1 metric is used in DocRE. The mathematical formula for these evaluation metrics can be found in Equation 2.1, 2.2, 2.3, 2.4, 2.5.

Table 2.6: Overview of the datasets used for document-level RE

Datasets	Paper	Year	Evaluation metrics	Relation types	Documents	Source
ChemProt	[111]	2011	Micro F1	14	1820	Biochemical
CDR	[112]	2016	F1	3116	1500	PubMed articles
TACRED	[40]	2017	F1	41	106,264	Newswire, Webtext
DocRED	[83]	2019	F1, Ign F1	96	5,053	Encyclopedic
GDA	[91]	2019	F1	2	30,192	Biomedical abstracts
DWIE	[117]	2019	F1	65	802	Deutsche Welle articles
HacRED	[115]	2021	F1	26	9231	Chinese text
ReDocRED	[116]	2022	F1, Ign F1	96	4053	Encyclopedic

Table 2.7: Overview of the datasets used for cross-document-level RE

Datasets	Paper	Year	Evaluation metrics	Relation types	Documents	Source
CodRED	[52]	2021	F1	276	5,882,234	General

$$\text{Precision} = \frac{\text{TruePostive}}{\text{TruePositive} + \text{FalsePositive}} \tag{2.1}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \tag{2.2}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

$$F1_{\text{Macro}} = \frac{1}{N} \sum_{i=1}^{N} F1_i \tag{2.4}$$

where N represents the total number of distinct classes or categories in the classification task.

$$F1_{\text{Micro}} = 2 \times \frac{\text{Total TruePositive}}{\text{Total TruePositive} + \text{Total FalsePositive} + \text{Total FalseNegative}} \tag{2.5}$$

2.7 Results

We discuss the results of sentence-level, document-level and cross-document-level RE with respect to rule based, supervised, unsupervised, distant-supervision and knowledge based methods. We compare various models over technique, the year in which the model was released, the dataset used for evaluation and results across different evaluation metrics. The detailed results are shown in Tables 2.8, 2.9, 2.10.

Table 2.8: The performance of different approaches for sentence-level relation extraction.

Technique	**Year**	**Model**	**Dataset**	**Source**	**F1 score**	**Precision**
Rule based method	1993	Autoslog	Terrorist event description	[119]	41.79	45
Rule based method	1993	Fastus	Newswire articles	[120]	-	44
Rule based method	2000	Snowball	Terrorist incidents	[121]	-	76
Rule based method	2013	-	DBpedia	[65]	68.1	75.5
Supervised learning (Feature based)	2010	UTD	SemEval-2010	[66]	82.19	82.25
Supervised learning (Feature based)	2004	MEMM	ACE	[67]	-	-
Supervised learning (Feature based)	2005	SVM based	ACE	[68]	74.7	84.8
Supervised learning (Kernel based)	2005	SVM based	ACE	[71]	52.50	65.50
Supervised learning (Kernel based)	2019	R-BERT	SemEval-2010 task 8	[39]	89.25	-
Supervised learning (Kernel based)	2010	SVM	ACE	[73]	-	-
Unsupervised method	2004	Clustering based	NYT	[75]	75 (avg)	76 (avg)
Distant supervision and knowledge based method	2018	RESIDE	Riedel	[78]		84.0
Knowledge based method	2020	BERT based	TACRED	[79]	68.8	70.3
Knowledge based method	2020	BERT based	Aggregated NewsRE	[79]	89.9	90.2
Knowledge based method	2021	RECON	Wikidata	[80]	87.23 (micro)	87.24
Knowledge based method	2021	RECON	NYT	[80]	-	87.5(P@10)

2.7.1 Comparative analysis of existing techniques and proposed approaches

Sentence-level relation extraction:

Table 2.9: The performance of different approaches for document-level relation extraction

Technique	Year	Model	Dataset/Domain	Reference	F1 score	Ign F1
Rule based method	2021	BiLSTM based model	DocRED	[86]	56.23	-
Rule based method	2019	Biomedical text	CDR	[47]	63.6	-
Rule based method	2019	Biomedical text	GDA	[47]	81.5	-
Rule based method	2020	GAIN-BERT	DocRED	[49]	62.76	60.31
Rule based method	2019	C-AGGCN	TACRED	[87]	69.0	-
Rule based method	2019	C-AGGCN	SemEval	[87]	85.7	-
Rule based method	2019	RGCN	CDR	[88]	58.6	-
Rule based method	2019	RGCN	CHR	[88]	87.5	-
Supervised method	2018	CAN	CDR	[89]	69.08	-
Graph based	2020	GAIN-BERT	DocRED	[49]	62.76	60.31
Graph based	2021	DRN-BERT	DocRED	[93]	61.37	59.15
Transformer based	2021	DocuNet-RoBERTa	DocRED	[94]	64.55	62.39
Transformer based	2020	RoBERTa-ATLOP	DocRED	[95]	63.40	61.39
Transformer based	2023	DREEAM	DocRED	[96]	67.53	65.47
Distant supervision	2020	BERT+D+P	DocRED	[97]	58.43	56.68
Knowledge based	2021	KB-Graph +kb-text	DWIE	[99]	52.1	-
Knowledge based	2022	ATLOP+KIRE	DocRED	[100]	61.39	59.35
Knowledge based	2021	KB-Graph +kb-text	DocRED	[99]	52.1	-
Knowledge based	2022	ATLOP+KIRE	DWIE	[100]	80.73	74.43
Knowledge based	2024	DocRE-CLip	DocRED	[51]	68.51	66.31
Knowledge based	2024	DocRE-CLip	DWIE	[51]	67.10	58.87
Knowledge based	2024	DocRE-CLip	ReDocRED	[51]	81.55	80.57

Table 2.10: The performance of different approaches for cross-document-level relation extraction

Technique	Year	Model	Dataset/Domain	Reference	F1 score
Distant supervision method	2006	RelEx	Wikipedia	[122]	-
Transformer & Graph-based model	2021	CodRED, closed (pipeline)	CodRED	[52]	32.29
Transformer & Graph-based model	2021	CodRED, closed (end-to-end)	CodRED	[52]	51.02
Transformer & Graph-based model	2021	CodRED, open (pipeline)	CodRED	[52]	28.70
Transformer & Graph-based model	2021	CodRED, open (end-to-end)	CodRED	[52]	45.06
Transformer & Graph-based model	2021	ECRIM, closed	CodRED	[102]	62.48

In this section, we compare various traditional models with our proposed approach, ReOnto (Refer Table 2.11). Traditional models such as Autoslog [119], Fastus [120], and Snowball [121] employ rule-based methods, which are effective for tasks with predefined, rigid target representations but often struggle with capturing complex textual nuances. Models like UTD [66], MEMM [67], and SVM-based approaches [68, 71] utilize supervised learning techniques, incorporating diverse lexical, syntactic, and semantic features. These models are robust in feature-rich environments but may suffer from overfitting or require extensive labeled training data. R-BERT [39] leverages the pre-trained BERT language model, achieving significant improvements in relation classification tasks but requiring substantial computational resources. Clustering-based methods [75] offer unsupervised relation discovery from large corpora, while RESIDE [78] and RECON [80] use neural networks and knowledge bases to enhance relation extraction performance. ReOnto, our proposed approach, addresses the task of Relation Extrac-

tion (RE) by combining neuro-symbolic knowledge with graph neural networks. It leverages publicly accessible ontologies as prior knowledge to identify sentential relations between entities. This hybrid approach balances precision and recall, making it adaptable and efficient for various applications, particularly in the biomedical domain where traditional techniques often yield unsatisfactory results. By integrating the strengths of both rule-based and statistical methods, ReOnto offers a robust solution for relation extraction tasks, ensuring high accuracy and adaptability across different domains.

Table 2.11: Comparison of various models for sentence-level relation extraction with proposed approach

Model	**Technique**	**Insights/Details**
Autoslog	Rule-based method	Automatically builds a domain-specific dictionary of concepts for information extraction
Fastus	Rule-based method	Uses finite-state automata for pattern matching
Snowball	Rule-based method	Generates patterns and extracts tuples from plain text
UTD	Supervised learning	Classification is performed using SVM classifiers
MEMM	Supervised learning	Employs maximum entropy models to combine diverse lexical, syntactic, and semantic features derived from the text
SVM-based	Supervised learning	Incorporates diverse lexical, syntactic, and semantic knowledge in feature-based relation extraction
SVM-based	Supervised learning	Employs the shortest path between the two entities using a dependency graph
R-BERT	Supervised learning	Leverages the pre-trained BERT language model and incorporates information from the target entities
Clustering-based	Unsupervised method	Groups multiple mentions together via clustering
RESIDE	Distant supervision and knowledge-based method	Improves distantly-supervised neural relation extraction using side information
RECON	Knowledge-based method	Uses GNN to learn representations of both the sentence and facts stored in a knowledge graph
ReOnto	Knowledge-based method	Ontology-based approach for relation extraction

Document-level relation extraction models:

Document-level relation extraction (DocRE) models have evolved significantly, leveraging diverse methodologies to address the inherent complexity of reasoning across multiple sentences. Traditional approaches, such as BiLSTM-based models, focus on heuristically selecting evidence sentences, achieving competitive performance despite their simplicity. Graph-based models, such as **GAIN-BERT** [49], **C-AGGCN** [87], and **RGCN** [88], construct intricate graphs to capture inter- and intra-sentence dependencies, enabling reasoning through multi-instance

learning or attention-guided mechanisms. Models like **CAN** [89] and **DRN-BERT** [93] emphasize dependency paths and discriminative reasoning frameworks, respectively, to enhance semantic understanding.

Meanwhile, **DocuNet-RoBERTa** [94] and **Roberta-ATLOP** [95] introduce novel techniques like adaptive thresholding and U-shaped segmentation to address multi-label and multi-entity challenges. **DREEAM** [96] and **BERT+D+P** [97] tackle evidence retrieval and distant supervision noise, improving memory efficiency and pre-training strategies. **KB-Graph+kb-text** [99] and **ATLOP+KIRE** [100] integrate external knowledge bases and coreference reasoning to enrich document-level representations.

In contrast, **DocRE-Clip** reframes DocRE as link prediction over a knowledge graph, combining entity context with document-derived logical reasoning. This approach enhances interpretability and reasoning quality, offering a unique perspective compared to graph-based and evidence-focused models. By evaluating on benchmark datasets, **DocRE-Clip** demonstrates its ability to balance contextual understanding and logical inference, setting itself apart in the DocRE landscape.

Models	**Technique**	**Insights/Details**
BiLSTM, Biomedical text, GAIN-BERT, C-AGGCN, RGCN	Rule-based	Uses predefined rules for decision making
CAN	Supervised	Requires labeled data for training
GAIN-BERT, DRN-BERT	Graph-based	Utilizes graph structures for data representation
DocuNet-RoBERTa, RoBERTa-ATLOP, DREEAM	Transformer-based	Uses attention mechanisms to process sequences
BERT+D+P	Distant supervision	Uses noisy labeled data for training
KB-Graph +kb-text, ATLOP+KIRE, DocRE-CLip	Knowledge-based	Incorporates external knowledge bases

Table 2.12: Comparison of Models for Document Level Relation Extraction with our proposed approach

Cross-Document-level Relation Extraction with our proposed approach: CrossDocRE signifies a substantial advancement in the domain of cross-document relation extraction, effectively addressing several limitations inherent in existing models such as RelEx [122], CodRED [52], and ECRIM [102]. Unlike RelEx [122], which primarily relies on distant supervision to reduce the necessity for manual annotation, CrossDocRE enhances its efficacy in cross-document contexts by integrating domain knowledge and providing interpretability. This integration allows CrossDocRE to offer a more nuanced understanding of relational contexts across multiple documents.

CodRED [52], which employs both pipeline and end-to-end methodologies utilizing transformer and graph-based models, introduces the concept of cross-document relation extraction. However, it lacks the integration of domain knowledge and interpretability features that are pivotal for comprehensive relation extraction. These are areas where CrossDocRE demonstrates

significant improvement. Similarly, ECRIM [102], which also combines transformer and graph-based techniques, does not account for domain knowledge, a critical gap that CrossDocRE effectively bridges.

By embedding domain knowledge, CrossDocRE not only enhances the interpretability of its predictions but also improves overall performance. This makes CrossDocRE a comprehensive and advanced approach for knowledge acquisition in diverse document settings, setting a new benchmark in the field of cross-document relation extraction. The model's ability to provide explanatory text for predicted relations further underscores its utility and effectiveness in practical applications.

Table 2.13: Comparison of models for Cross-Document-level Relation Extraction with our proposed approach

Technique	**Model**	**Insights/Details**
Distant supervision method	RelEx	Uses distant supervision for training, reducing the need for manual annotation
Transformer & Graph based model	CodRED, closed (pipeline)	Combines transformer and graph-based methods in a pipeline approach
Transformer & Graph based model	CodRED, closed (end-to-end)	Integrates transformer and graph-based methods in an end-to-end framework
Transformer & Graph based model	CodRED, open (pipeline)	Utilizes a pipeline approach with transformer and graph-based techniques
Transformer & Graph based model	CodRED, closed (end-to-end)	End-to-end integration of transformer and graph-based methods
Transformer & Graph based model	ECRIM, closed	Closed system combining transformer and graph-based techniques
Transformer & Graph based model	CrossDocRE, open and closed	Closed system combining transformer and graph-based techniques

2.7.2 Optimal RE Models

Several models have been proposed for each relation extraction dataset. We examine the most effective models put forward for each dataset (refer Tables 2.14, 2.15).

REBEL: For the `CoNLL04 dataset`, REBEL [123] is the most promising model so far in terms of the F1 score. This model uses autoregressive seq2seq models that perform end-to-end relation extraction for more than 200 different types. The flexibility of the model is fine-tuned by an array of relation extraction and relation classification benchmarks. The model was released in the year 2021, and the macro F1 score is 76.6.

PL-Marker: For the `ACE 2004` and `ACE 2005` datasets, the most optimal model is [124]. It considers the interrelation between spans. In this work, the authors propose a novel span repre-

sentation approach, named Packed Levitated Markers (PL-Marker), to consider the interrelation between the spans (pairs) by strategically packing the markers in the encoder. In particular, it uses a neighbourhood-oriented packing strategy, which considers the neighbour spans to model the entity boundary information.

UniRel and SPN: For the `NYT` dataset, two models stand out as the most effective. The first is UniRel, which excels at the relation extraction task, and the second is SPM, which performs well in joint entity and relation extraction tasks. UniRel model is a unified representation and interaction for joint relational triple extraction [125]. UniRel unifies the representations of entities and relations by jointly encoding them in a concatenated natural language sequence, and unify the modeling of interactions with a proposed interaction map, which is built upon the off-the-shelf self-attention mechanism in any transformer block.

For the pipeline task of joint entity and relation extraction, a set prediction networks (SPN) model was proposed [126]. Unlike traditional methods that generate triples sequentially, the proposed SPN model outputs the final set of triples all at once.

SP: The most optimal model for the `SemEval-2010 task-8` dataset was proposed in the year 2020 and it is called SP. They present a span-prediction based system for relation classification and evaluate its performance by comparing it with the embedding based systems [127]. They demonstrate that the supervised SP objective works significantly better than the standard classification based objective.

ReOnto: `ADE` is the adverse drug effect dataset containing biomedical text and labels, and the most optimal model for this dataset proposed thus far is ReOnto [15]. ReOnto employs a graph neural network to acquire the sentence representation and leverages public ontologies as prior knowledge to identify the sentential relation between two entities. ReOnto was released in the year 2023 and has an F1 score of 90.6. More details of this model are given in Chapter 4.

SciBERT: The most suitable model for the `Chemprot` dataset is SciBERT [128], a pretrained language model based on BERT designed to address the shortage of high-quality, large-scale labelled scientific data. SciBERT leverages unsupervised pretraining on a large multi domain corpus of scientific publications to improve the performance on downstream scientific NLP tasks.

SAIS: SAIS: The SAIS model has achieved the highest F1 score to date on the `CDR` dataset [129]. It explicitly teaches the model to capture relevant contexts and entity types by supervising and augmenting intermediate steps for relation extraction (RE). Based on a broad spectrum of carefully designed tasks, the SAIS method extracts relations of better quality due to more effective supervision and retrieves the corresponding supporting evidence more accurately, thereby en-

hancing interpretability.

RAG4RE: RAG4RE model has achieved the highest F1 score on the TACRED dataset [130]. RAG4RE was proposed to incorporate Large Language Models (LLM) for the relation extraction task by addressing the hallucination issue of LLM. RAG4RE method used LLMs such as Flan T5, Llama2 and Mistral for testing the validity of the method.

DocRE-CLip: DocRED is the most popular dataset for document-level relation extraction. A knowledge based method, DocRE-CLip, is the best model so far on this dataset, which is released in year 2024. DocRE-CLip reformulates the document-level relation extraction task as a link prediction task and incorporates domain knowledge of entities (entity context and entity connecting path) in the form of triples. DocRE-CLip also aggregates the reasoning module, which implements intra sentence reasoning, logical reasoning and co-reference reasoning with link prediction task. DocRE-CLip works best for the DWIE and ReDocRED datasets.

SAISORE+CR+ET-SciBERT and seq2rel: The GDA dataset includes labelled genes, diseases, and gene-disease associations obtained through distant supervision. Similar to the CDA dataset, SAISORE+CR+ET-SciBERT is the optimal model for the GDA dataset in the relation extraction task. For the joint entity and relation extraction, the seq2rel model reported the highest F1 score [131]. The model is implemented using a sequence-to-sequence approach, seq2rel, that can learn the subtasks of DocRE such as entity extraction, co-reference resolution and relation extraction.

KD-Rb-l: The HacRED dataset is designed to be more challenging by including hard cases. Few methods evaluate their models on this dataset.The KD-Rb-l method is the optimal model among the few, with an F1 score of 67.28. KD-Rb-l is a semi-supervised framework for DocRE with three components. Firstly, they use an axial attention module for learning the interdependency among entity pairs, which improves the performance on two hop relations. Secondly, they propose an adaptive focal loss to tackle the class imbalance problem of DocRE. Lastly, they use knowledge distillation to overcome the differences between human annotated data and distantly supervised data. This model was released in the year 2022.

KXDocRE: CodRED is the first and only dataset available for cross-document relation extraction work. It is constructed using Wikipedia. Knowledge based method, named KXDocRE, has the highest F1 score reported so far on this dataset. It considers the entity context and connecting path context from Wikidata to feed into the transformer based model.

Table 2.14: Description of the best model for different sentence-level datasets

Datasets	Task	Model	Model released year	Source	F1 score
CoNLL04	RE	REBEL	2021	[123]	76.65 (macro)
ACE2004	RE	PL-Marker	2021	[124]	66.5 (micro)
ACE2005	RE	PL-Marker	2021	[124]	73.0 (micro)
NYT	RE	UniRel	2022	[125]	93.7
NYT	Joint entity and RE	SPN	2020	[126]	92.5
SemEval-2010 Task-8	RE	SP	2020	[127]	91.9
ADE	RE	ReOnto	2023	[15]	90.6

Table 2.15: Description of the best model for different document-level and cross-document-level datasets

Datasets	Task	Model	Model released year	Source	F1 score
ChemProt	Joint entity and DocRE	aimped	2023	-	85 (macro)
ChemProt	DocRE	SciBert	2019	[128]	83.64
CDR	DocRE	SAISORE+CR+ET-SciBERT	2022	[129]	79
TACRED	DocRE	RAG4RE	2024	[130]	86.6
DocRED	DocRE	DocRE-CLiP	2024	[116]	68.51
GDA	DocRE	SAISORE+CR+ET-SciBERT	2021	[129]	87.1
GDA	Joint entity and DocRE	seq2Rel	2022	[131]	55.2
DWIE	DocRE	REXEL	2024	[132]	65.8(hard)
HacRED	DocRE	KD-Rb-l	2022	[133]	67.28
ReDocRED	DocRE	DocRE-CLiP	2024	[51]	81.55
CodRED	CrossDocRE (closed setting)	KXDocRE	2024	[134]	66.3
CodRED	CrossDocRE (open setting)	KXDocRE	2024	[134]	57.12

2.8 Large Language Models For Relation extraction

Large language models (LLMs) use deep learning techniques and massive datasets to understand, summarize, generate and predict new content. LLMs are trained on huge data to learn patterns and entity relationships in the language. LLMs can be used in various tasks such as translating languages, analyzing sentiments, question answering, grammar correction and others. However, LLMs pose several challenges such as hallucination, and incorrect results. Large language models can also be used for relation extraction tasks. The underlying architecture is a set of neural networks that consist of an encoder and a decoder with self-attention. The transformer extract meanings from a sequence of text and understand the relationships between words and phrases in it. Unlike earlier recurrent neural networks (RNN) that sequentially process inputs, transformers process entire sequences in parallel. This allows the data scientists to use GPUs for training transformer-based LLMs, significantly reducing the training time. Transformer neural network architecture allows the use of very large models, often with hundreds of billions of parameters.

2.8.1 Key components of LLMs:

We briefly discuss the key components of the LLMs that are required to explain the LLM architecture. The key components of LLMs are tokenizers, embedding layer, feed-forward layer, recurrent layer, and attention module to process the text and generate output context (refer Figure 2.4).

Tokenization: Tokenization is the basic step in large language models. It is the process of splitting the text into smaller units or tokens.

Embedding: The process of creating embeddings from text takes place in embedding layer, which captures the semantic and syntactic meaning of the input, so that the model can understand context.

Feed forward layer: The feed-forward layer (FFN) of a large language model is made up of multiple fully connected layers that transform the input embeddings. These layers enable the model to glean higher-level abstractions i.e., to understand the user's intent with the text input.

Recurrent layer: The recurrent layer interprets the words in the text in sequence. It captures the relationship between words in a sentence.

Attention module: The attention mechanism enables a language model to focus on single parts of the text that are relevant to the task. This layer allows the model to generate the most accurate outputs.

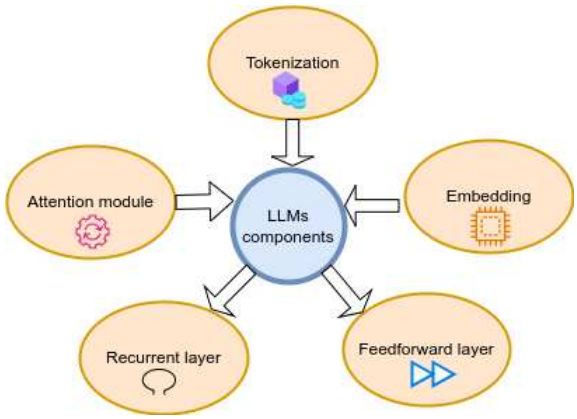

Figure 2.4: Key components of LLMs

2.8.2 LLM Architecture

The text is tokenized into smaller units called tokens, and each token is embedded into a continuous vector representation. Following this, positional encoding is added to the input embeddings to provide details about the positions of the token. Transformers do not naturally encode

the order of the tokens. So, position encoding ensures that the position of the token is saved. Depending on the neural network, the encoder analyses the text and creates a number of hidden states that protect the context and meaning of text data. There are two key components in encoder layer a) self-attention layer which enables the model to weigh the importance of different tokens in the input sequence by computing attentions scores and b) after the self attention layer, a feed forward layer is applied to each token independently. This network includes fully connected layers with non-linear activation functions, allowing the model to capture complex interactions between tokens.

The decoder receives as input its own predicted output word at each time-step. The input to the decoder is also augmented by positional encoding. The augmented decoder input is fed into the three sub layers comprising the decoder block. Masking is applied in the first sub layer to stop the decoder from attending to the succeeding words. At the second sub layer, the decoder also receives the output of the encoder, which now allows the decoder to attend to all the words in the input sequence. The output of the decoder finally passes through a fully connected layer, followed by a softmax layer, to generate a prediction for the next word of the output sequence.

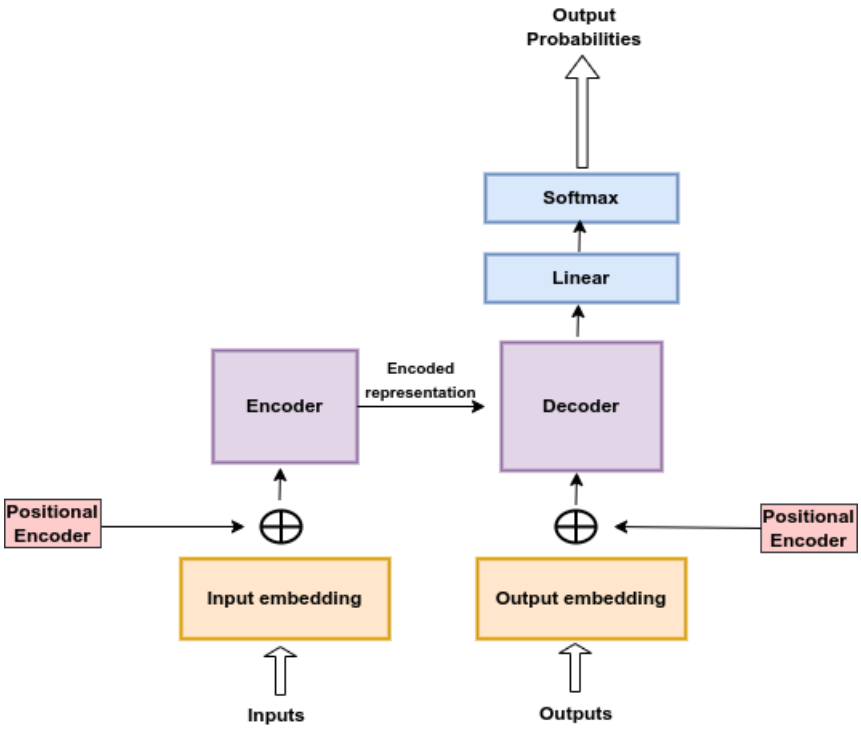

Figure 2.5: Underlying architecture diagram of transformer

2.8.3 Examples of Large Language Models

GPT-4: GPT stands for Generative Pre-trained Transformer. GPT-4 is the latest release from OpenAI. It was launched on March 14, 2023. GPT-4 offers multi-modal capabilities. Official documentation for GPT-4 is not yet available, but it is estimated to have parameters in the trillions and an unknown maximum sequence length. As the most advanced model in the GPT series to date, GPT-4 can handle complex tasks like translating the sentence "The lawyer argued that the defendant's actions were not a proximate cause of the plaintiff's injuries, but rather an intervening cause." to French as "L'avocat a soutenu que les actions du défendeur n'étaient pas une cause proche des blessures du plaignant, mais plutôt une cause intervenante". However, generating such accurate translations may take longer than with previous models [135]

GPT-3 GPT-3, released in June 2020, boasts 20 billion parameters and a maximum sequence length of 4096 (Brown et al., 2020). It represents a significant advancement in the field of large language models. This model has demonstrated remarkable capabilities in few-shot learning, allowing it to perform a wide range of tasks with minimal task-specific training. Its ability to generate coherent and contextually relevant text has made it a benchmark for evaluating the performance of other large language models.

Gemini: Google launched the Gemini series and released various versions such as Ultra, Pro and Nano with a focus on multi-modality (text, audio, image, video support). Gemini improved on the state-of-the-art LLMs and performed better than ChatGPT and GPT4-vision models in many of the tests, including language comprehension, contextual understanding, problem-solving, creative writing, and technical analysis. It was released in December, 2023.

LlaMA: LlaMa is a family of large language models released by Meta AI. It was released in April, 2024. It is a collection of foundation language models ranging from 7B to 65B parameters. The LlaMa model is trained on trillions of tokens and shows that it is possible to train state-of-the-art models using publicly available datasets exclusively, without the need for proprietary and inaccessible datasets.

PaLM: PaLM (Pathways Language Model) is a 540 billion parameter transformer-based large language model developed by Google AI. Researchers also trained smaller versions of PaLM, 8 and 62 billion parameter models, to test the effects of model scale.

Details of some popular large language models are given in Table 2.16.

Table 2.16: Descriptions of some popular large language models

Series	Source	Model	Size	Open Source	Instruction Tuning
BART	[136]	BART	140M (base)	yes	-
T5	[137]	T5	60M, 220M (base), 770M (large), 3B, 11B	yes	-
T5	[138]	mT5	300M, 580M (base), 1.2B (large), 3.7B, 13B	yes	-
T5	[33]	Flan-T5	80M, 250M (base), 780M (large), 3B, 11B	-	-
GPT	[139]	GPT-2	117M, 345M, 762M 1.5B	yes	yes
GPT	[140]	GPT-3	6B	yes	-
GPT	[141]	GPT-J	6B	yes	-
GPT	[142]	Code-davinci-002	-	-	yes
GPT	[142]	Text-davinci-002	-	-	yes
GPT	[142]	Text-davinci-003	-	-	yes
GPT	[143]	GPT-3.5-turbo series	-	-	yes
GPT	[135]	GPT-4 series	-	-	yes
LLaMA	[144]	LLaMA	7B, 13B, 33B, 65B	yes	-
LLaMA	[145]	Alpaca	7B, 13B	yes	yes
LLaMA	[146]	Vicuna	7B, 13B	yes	yes
LLaMA	[144]	LLaMA2	7B, 13B, 70B	yes	-
LLaMA	[144]	LLaMA2-chat	7B, 13B, 70B	yes	yes
LLaMA	[147]	Code-LLaMA	7B, 13B, 34B	yes	-

2.8.4 Relation extraction using LLM

LLMs often perform poorly on RE tasks due to the low incidence of RE in instruction-tuning datasets [148]. To address these concerns, QA4RE [149] introduced a framework that enhances LLMs' performance by aligning RE tasks with QA tasks. In a similar case, SUMASK (summarize and ask) was proposed, which attempts to solve RE as zero-shot relation extractors using prompt techniques such as chain-of-thought to improve zero-shot RE. To address the low relevance between entity and relation, the inability to explain input-label mappings, and the limitations of the pipeline approach, such as error propagation and the dependence on intermediate task performance, GPT-RE [150] integrates task-aware representations and enriches demonstrations with reasoning logic. Additionally, to mitigate these issues associated with pipeline approaches, the end-to-end language generation model, REBEL [123], was proposed, which employs autoregressive seq2seq models for an end-to-end language generation task.

To strengthen the discriminating power of contrastive learning and increase the diversity of positive pairs, AugURE was proposed which uses both within-sentence pairs augmentation and augmentation through cross sentence pairs [151]. RELA studies the effect of relation names and their synonyms, textual semantics and the correlation among them [152]. RELA is a Seq2Seq model with automatic label augmentation for RE. In another study, CoT-ER (chain-of-thought with explicit evidence reasoning) was proposed, which is used for few-shot relation extraction tasks using large language models [153]. In particular, CoT-ER first induces large language models to generate evidence using task-specific and concept-level knowledge. Then the evidence is explicitly incorporated into chain-of-thought prompting for relation extraction. With the strong synthetic data generation power of LLMs, REPaL was introduced to utilize LLMs for generating initial seed instances based on relation definitions. It fine-tunes a bi-directional small language model using these initial seeds to learn the relations from the target domain [154]. They enhance pattern coverage and mitigate bias resulting from the limited number of initial seeds by incorporating feedback acquired from SLM's predictions on unlabelled corpora. Recent studies have introduced MICRE (Meta In-Context Learning of LLMs for Relation Extraction), a novel framework for zero and few-shot relation extraction. This framework involves tuning an LLM to perform in-context learning on a diverse collection of RE datasets [155]. To tackle the challenges of poor performance in a supervised setting, a method called recall-retrieve-reason RE was proposed, which synergizes LLMs with retrieval corpora (training examples) to enable relevant retrieving and reliable in-context reasoning. Furthermore, they used consistent ontological knowledge from training datasets to let LLMs generate relevant entity pairs grounded by retrieval corpora as valid queries. These entity pairs are then used to retrieve relevant training examples from the retrieval corpora as demonstrations for LLMs to conduct better in-context learning via instruction tuning [156]. To handle hallucinations, RAG4RE (Retrieved-Augmented Generation-based Relation Extraction) work was proposed, which offers a pathway to enhance the performance of relation extraction tasks [130]. Details of these models are given in Table 2.17.

Table 2.17: Description of best performing LLMs on sentence-level RE datasets

Model	Year	Technique	Backbone	Datasets used
REBEL	2021	Supervised fine tuning	BART-large	NYT, ADE
QA4RE	2023	Zero shot	Text-davinci-003	TACRED, Re-TACRED, TEACREV, SemEval
SUMASK	2023	Zero shot	GPT-3.5-turbo-0301	TACRED, Re-TACRED, TECREV
GPT-RE	2023	In context learning	Text-davinci-003	TACRED, SemEval
AugURE	2023	contrastive learning	ChatGPT	NYT-FB, TACRED
RELA	2023	Seq2Seq	BART	TACRED, SemEval, Google RE, sciERC
REPaL	2024	In context learning	GPT-4	DefOn-FewRel, DefOn-ReTACRED

2.9 Summary

Our study focuses on the overview of RE, categorized by the granularity of text. RE is a promising field and has been growing since many decades. Due to the wide range of applications and challenges associated with RE, numerous methods have been proposed to tackle these tasks from different perspectives. We have briefly discussed rule based, supervised, unsupervised, distant supervision and knowledge based methods with respect to the granularity of the text.

CHAPTER 3

Background

3.1 Symbolic AI

Symbolic AI is a branch of artificial intelligence that relies on the use of explicit symbols and rules to represent knowledge and perform reasoning. This approach is based on the idea that human thought can be represented through the manipulation of symbols and logical operations. Symbolic AI systems use predefined rules and symbols to represent knowledge. These rules are often written in formal languages and are used to perform logical reasoning. Knowledge in symbolic AI is typically represented using structures such as semantic networks, frames, and ontologies. In this thesis, we use ontologies and knowledge graphs. So they are discussed in more detail in the following sections.

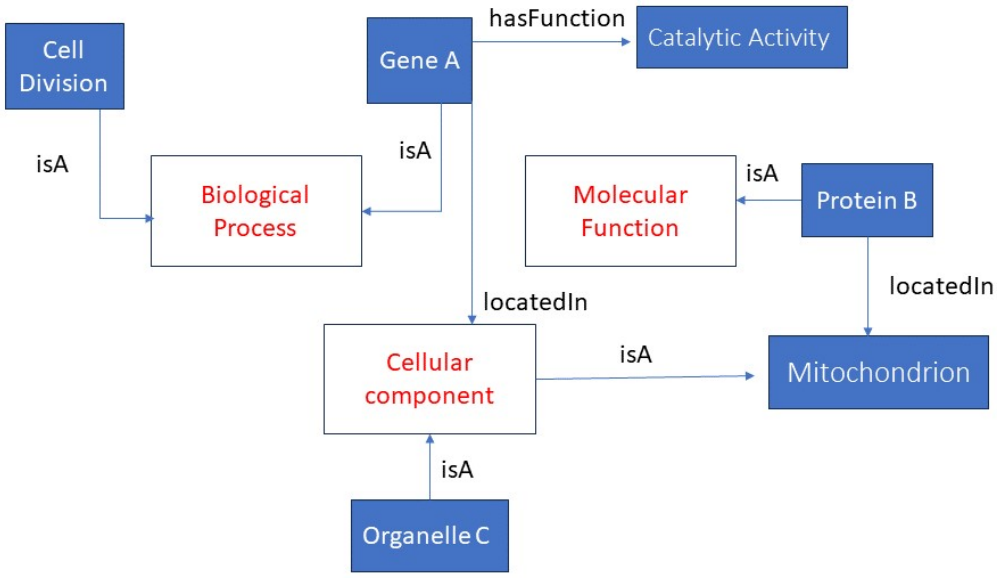

Figure 3.1: Subgraph of Gene ontology

3.1.1 Ontologies

An ontology is a knowledge representation that defines a set of concepts and relationships between those concepts [157]. Ontologies enhance the capability of computer systems to process, understand, and generate human-like understanding of data and language across multiple domains. To understand the major components of ontology, consider an example of one of the

most popular ontologies in the field of knowledge representation, i.e., Gene ontology [158] (refer Fig 3.1). Gene ontology is a structured, controlled vocabulary that describes gene and gene product attributes in any organism. A few key concepts of ontologies are as follows.

Classes: Classes describe the concepts in the domain. For Gene ontology, some of the classes are as follows. 1) Biological Process: represents a biological objective to which the gene or gene product contributes. 2) Molecular Function: describes the elemental activities of a gene product at the molecular level. 3) Cellular Component: refers to the parts of a cell or its extracellular environment.

Relationships: Relationship (also known as a property or relation) is a connection or association between concepts or entities. There are two types of relationships (properties) in the ontology - object property and data property. Object properties connect two individuals, while data properties connect individuals with numeric or string constants. Here are some examples of object and data properties from Gene ontology.

- is_a: Represents a subclass relationship. Example: `Cell Division` is_a `Biological Process`
- has_function: It is an object property which relates a gene product to its molecular function. Example: `Gene A` has_function `Catalytic Activity`
- located_in: It is an object property which relates a gene product to its cellular component. Example: `Protein B` located_in `Mitochondrion`
- has_sequence_length: It is a data property which represents sequence length. Example: `Protein B` has_sequence_length 350
- has_molecular_weight: It is a data property which represents the property molecular weight. Example: `Protein C` has_molecular_weight 55.5.

Individuals: These are the instances of the classes. For example, Gene A, Protein B, and Organelle C are individuals of Biological process, Molecular Function and Cellular component class, respectively.

Axioms: Axioms refer to statements that capture the relationship between the classes. Axioms can describe various aspects, including class hierarchies, relationships between classes or individuals, property characteristics, domain and range restrictions, cardinality constraints, and logical implications [159]. Some examples of these axioms from gene ontology are as follows.

- Subsumption Axioms: `Biological Process` is a subclass of `Process`. Molecular Function is a subclass of `Function`. `Cellular Component` is a subclass of `Component`.

- Disjointness Axioms: `Biological Process` is disjoint from `Molecular Function`. `Biological Process` is disjoint from `Cellular Component`. `Molecular Function` is disjoint from `Cellular Component`.
- Range Restriction Axiom: This construct is used to define the range of the property. For example, `has_molecular_weight rdfs:range xsd:float` expresses that the property has_molecular_weight will always have a float value. This axiom indicates that any value assigned to the has_molecular_weight property must be of the type xsd:float.

Applications: Ontologies play a crucial role in structuring and organizing knowledge across various domains. In knowledge representation, ontology serves as the backbone for modelling domain-specific knowledge in a structured and formal manner [160]. They provide a shared vocabulary for entities within a domain and define the relationships between these entities. This structured representation facilitates reasoning about the entities and their relationships, enabling systems to make inferences, support decision-making processes, and enhance knowledge discovery. For example, in healthcare, an ontology can represent the complex relationships between symptoms, diseases, and treatments, aiding in diagnosis and research [161]. In natural language processing (NLP), an ontology can contribute to the understanding and generation of natural language by providing structured knowledge about the world that can be used to infer context and meaning [162]. They help in tasks such as semantic search, text classification, information extraction, and question answering. By mapping parts of speech or phrases to entities and concepts in an ontology, NLP systems can better grasp the semantics of text, leading to more accurate interpretation of user queries, sentiment analysis, and content summarizing. Beyond these areas, ontologies are also applied in fields such as bioinformatics for gene classification, e-commerce for product categorization, and artificial intelligence for enabling machines to understand complex human concepts [163, 164, 160].

Related Standards and Tools: We discuss some of the World Wide Web Consortium (W3C) standards and tools related to ontologies here.

Web Ontology Language (OWL): OWL [165] is a computational logic-based language developed by the World Wide Web Consortium (W3C) for defining and instantiating ontologies. It is designed to represent rich and complex knowledge about things, groups of things, and relations between things. It is primarily serialized in RDF/XML format.

OWL is a key technology for the Semantic Web, which is an extension of the World Wide Web that enables people to share content beyond the boundaries of applications and websites. OWL allows for the creation, sharing, and processing of extensive and sophisticated ontologies on the World Wide Web. There are different variants of OWL, with OWL 2 being the latest, offering more expressive power for defining and classifying information.

Resource Description Framework (RDF): RDF is a standard model for data interchange on the Web [166]. It enhances the Web's linking structure by using URIs to name the relationships between entities as well as the entities themselves. RDF is often used alongside OWL to create a comprehensive framework for ontology modeling.

SPARQL (SPARQL Protocol and RDF Query Language): SPARQL is a semantic query language designed for databases that store data in the RDF format [167]. It allows users to retrieve and manipulate data by querying ontologies and extracting information based on specific patterns.

Protégé: Protégé [168] is a free, open-source ontology editor and a knowledge management system. It was developed by the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine. Protégé provides a graphical user interface for designing and editing ontologies. It has tools for structuring, managing, and visualizing complex ontologies. It also supports the development of custom plugins to extend its functionality, making it highly adaptable to different project needs. Protégé is widely used in the academic, corporate, and government sectors, making it one of the most popular ontology editing tools.

3.1.2 Knowledge Graphs

Knowledge graphs are data structures that represent knowledge in a graph format, comprising entities (nodes) that denote objects, concepts, events, or people, and relationships (edges) that connect these entities to illustrate how they are related [169]. These graphs are designed to facilitate the understanding, sharing, and analysis of knowledge in a way that is both machine-readable and meaningful to humans, enabling advanced applications in search, data integration, artificial intelligence, and beyond. Figure 3.2 shows a simple example of the knowledge graph. It shows relationships between entities such as Bonn and Fischer.

Applications: Knowledge graphs have a wide range of applications across various domains.

- Search Engines is one of the applications of KG. KGs enhance search results by understanding the context and relationships between different entities and concepts [170].
- Recommendation Systems utilize Knowledge Graphs (KGs) to provide personalized recommendations by analyzing the relationships between users, products, and preferences. Additionally, KGs facilitate the integration of diverse data sources by mapping them onto a unified graph structure, making it easier to identify relationships and insights across datasets [171].
- Knowledge graphs can enhance chatbots by providing them with a structured understanding of various entities and their relationships. This allows chatbots to deliver more accurate and contextually relevant responses, improving user interaction and satisfaction [172].

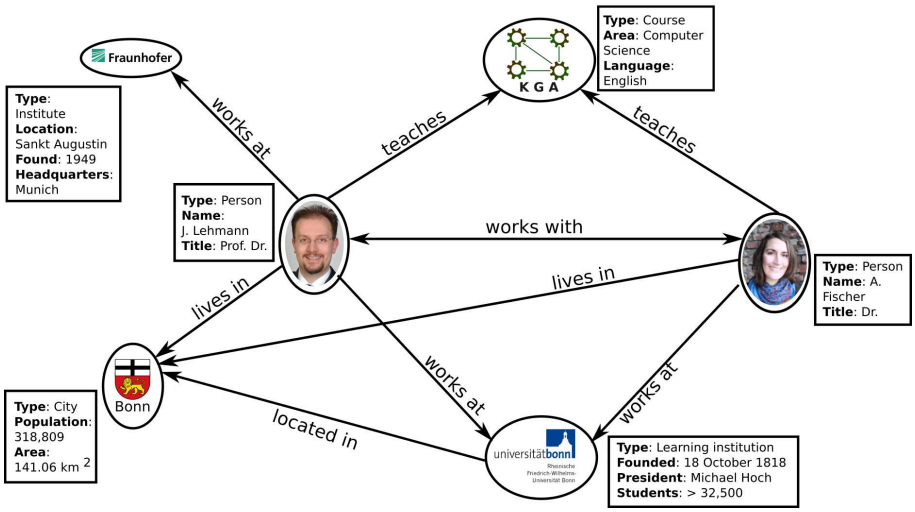

Figure 3.2: Example of knowledge graph [1]

- Knowledge graphs provide personalized recommendations by analyzing the relationships between users, products, and preferences. By understanding the connections and context, recommendation systems can suggest items that are more likely to be of interest to the user [173].
- Knowledge graphs power question answering systems by enabling them to understand the context and relationships between different entities. This allows the systems to provide more accurate and detailed answers to user queries, leveraging the rich, structured data within the knowledge graph [174].

Knowledge Graph Representation

Knowledge Graphs can be represented using different formats, with RDF (Resource Description Framework) and property graphs being the most common.

3.1.3 RDF (Resource Description Framework)

RDF is a standard model for data interchange on the web. It uses triples (subject-predicate-object) to represent data, making it easy to merge data from different sources. RDF is particularly useful for representing metadata and linking data across different domains. For example, RDF is used in the Semantic Web to enable data interoperability and integration [175].

3.1.4 Property Graphs

Property graphs are another popular representation format for Knowledge Graphs (KGs). They consist of nodes, edges, and properties, where nodes represent entities, edges represent relationships, and properties provide additional information about nodes and edges. For example, in a

property graph representing a social network, a node might represent a person, an edge could represent a "friendship" relationship, and properties might include attributes like age or location. Property graphs are widely used in graph databases like Neo4j, which support complex queries and analytics on graph data [175]. By understanding these representation formats, we can better appreciate how Knowledge Graphs are structured and utilized in various applications. This ties back to our earlier discussion on RDF, highlighting its role in the broader context of Knowledge Graphs. For instance, an RDF graph might represent the same social network using triples, such as <PersonA> <isFriendsWith> <PersonB>, where each triple consists of a subject, predicate, and object.

3.2 Sub-symbolic/Connectionist AI

Sub-symbolic or Connectionist AI refers to a branch of Artificial Intelligence that focuses on developing systems that learn and process information in ways inspired by the human brain. Unlike symbolic AI, which relies on explicit rules and symbols to represent knowledge, sub-symbolic AI uses neural networks and other connectionist models to process data in a more distributed and parallel manner. The primary structure used in sub-symbolic AI is the artificial neural network, which consists of interconnected nodes (neurons) that work together to process information. Sub-symbolic AI systems learn from data and are typically trained using large datasets. They learn patterns and features directly from the data rather than being programmed with explicit rules. An overview of some of the artificial neural networks relevant to our work is below.

GNN (Graph Neural Network):

Graph Neural Networks (GNNs) are a class of neural networks designed to work directly with graph-structured data [176]. Unlike traditional neural networks that operate on fixed-size input data (like images or sequences), GNNs can handle data represented as graphs, which consist of nodes (vertices) and edges (connections between nodes). This makes GNNs particularly useful for tasks where relationships and interactions between entities are crucial, such as social network analysis, molecular chemistry, and recommendation systems.

Architecture: The architecture of a GNN consists of multiple layers, each responsible for aggregating and updating information from neighbouring nodes. The core idea behind GNNs is the "message-passing" paradigm, where information is exchanged between nodes during the training process (Figure 3.3). At each layer, the GNN performs two fundamental steps.

Message passing: In this step, every node aggregates information from its neighbouring nodes, which is then transformed into an informative message. The message typically consists of

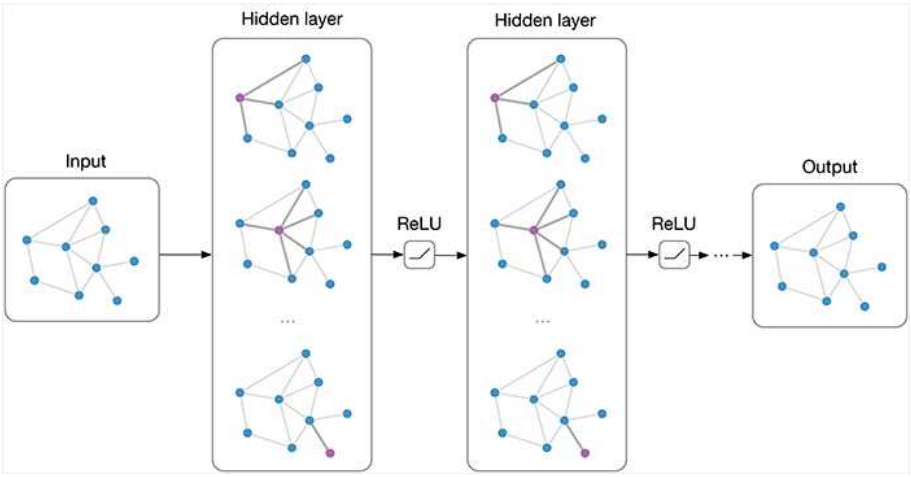

Figure 3.3: Basic architecture of GNN [2]

information from the node's features and the features of its neighbours.

Node update: The node utilizes the received messages to update its internal representation or embedding. This step enables nodes to incorporate information from their local neighbourhood and refine their representation. By iteratively repeating the message passing and node update steps, the GNN enables information to propagate across the entire graph, allowing nodes to learn and refine their embeddings collectively.

Challenges: There are several challenges in building and using GNNs, such as a) Scalability: GNNs can be computationally intensive, especially for large graphs. Efficient algorithms and hardware acceleration are needed to handle large-scale data. b) Over-smoothing: As the number of message-passing layers increases, node representations can become too similar, leading to a loss of discriminative power. c) Expressive power: Some GNN architectures may not be able to distinguish between certain graph structures, limiting their expressive power. d) Dynamic graphs: Handling graphs that change over time (dynamic graphs) presents additional challenges, as the model needs to adapt to evolving structures. e) Interpretability: Understanding how GNNs make predictions is difficult due to their complex nature. Interpretability is crucial for applications where transparency is important, such as healthcare and finance. f) Data Sparsity: Real-world graphs often have sparse connections, making learning meaningful representations challenging. g) Robustness: GNNs can be sensitive to adversarial attacks, where small perturbations in the graph structure can significantly affect the model's performance.

GNNs come in various forms, each tailored to handle specific types of graph-structured data. Some common types of GNNs are as follows.

Graph Convolutional Networks (GCNs): GCNs are one of the earliest and most widely used

GNN variants. They leverage graph convolutions to aggregate and update node representations based on their local neighbourhood [177].

GraphSAGE: GraphSAGE (Graph Sample and Aggregate) is another popular GNN architecture. It utilizes a sampling strategy to aggregate and update node embeddings, allowing for scalability in large graphs [178].

Graph Attention Networks (GATs): GATs introduce attention mechanisms into GNNs, enabling nodes to selectively attend to relevant neighbours while performing message passing and aggregation [179].

Graph Isomorphism Networks (GINs): GINs focus on capturing the structural information of graphs by applying permutation-invariant functions during the message passing and node update steps [180].

Applications of Graph neural networks (GNNs)

Graph Neural Networks (GNNs) have become increasingly popular due to their ability to model and analyse graph-structured data. Here are some key applications of GNNs:

- GNNs are used to analyse social networks by modeling relationships and interactions between users. They can be used for tasks such as community detection, link prediction, and node classification [181].
- GNNs are applied in molecular chemistry to predict molecular properties, understand molecular structures, and assist in drug discovery by modeling molecules as graphs where atoms are nodes and bonds are edges [182].
- GNNs can enhance recommendation systems by modeling user-item interactions as a bipartite graph, allowing for more accurate and personalized recommendations [183].
- GNNs are used to model and predict traffic flow and transportation networks by representing road networks as graphs, which helps in optimizing routes and managing traffic congestion [184].
- GNNs are employed to enhance knowledge graphs by performing tasks such as entity classification, link prediction, and reasoning over the graph to infer new knowledge [185].

In one of our proposed method (ReOnto, Chapter 4), we have employed Graph Neural Networks (GNNs) for text classification due to their capability to effectively capture complex relationships and dependencies within structured data.

Embeddings

Embeddings are a type of word representation that allows words with similar meanings to have a similar representation [186]. They are a distributed representation for text that is perhaps one of the key breakthroughs for the impressive performance of deep learning methods on challenging

natural language processing problems. The main idea behind embeddings is to map words or phrases from a vocabulary to vectors of real numbers in a low-dimensional space. This transformation helps in capturing semantic or syntactic similarity, relation with other words, etc.

How Embeddings Work

1. **Initialization**: Words are initially represented as one-hot encoded vectors. For example, in a vocabulary of 10,000 words, each word is represented by a vector of size 10,000 with a 1 in the position corresponding to that word and 0s elsewhere.
2. **Training**: During the training process, these one-hot vectors are transformed into dense vectors of much lower dimension (e.g., 100, 200, or 300 dimensions). This is typically done using neural network models that learn to predict the context of words. The weights of the neural network, which are updated during training, eventually become the word embeddings.
3. **Contextual Information**: Modern embedding techniques, such as those used in BERT, consider the context in which a word appears. This means that the same word can have different embeddings depending on its context, capturing nuances in meaning.
4. **Usage**: Once trained, these embeddings can be used in various NLP tasks such as sentiment analysis, machine translation, named entity recognition, and more. They provide a rich and dense representation of words that can significantly improve the performance of these tasks.

BERT (Bidirectional Encoder Representations from Transformers:)

BERT, developed by Google, is a transformer-based model that has revolutionized the field of NLP [187]. Unlike previous models that read text sequentially (either left-to-right or right-to-left), BERT reads the entire sequence of words at once. This allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

Key Features of BERT

- **Bidirectional Training**: BERT is trained bidirectionally, meaning it considers the context from both directions (left and right) for each word. This helps in understanding the full context in which a word is used.
- **Transformer Architecture**: BERT is based on the transformer architecture, which uses self-attention mechanisms to weigh the importance of input elements with respect to each other.
- **Pre-training and Fine-tuning**: BERT is pre-trained on a large corpus of text data and then fine-tuned for specific tasks. This two-step process allows BERT to leverage vast amounts of data and adapt to various NLP tasks with minimal task-specific data.

- **Masked Language Model (MLM)**: During pre-training, some of the words in the input are randomly masked, and the model is trained to predict these masked words. This helps the model learn deep bidirectional representations.

In each of our proposed methods (ReOnto in Chapter 4, DocREClip in Chapter 5, and KX-DocRE in Chapter 6), we utilized BERT to convert text into vector space representations.

BioBERT

BioBERT is a variant of BERT that is specifically designed for biomedical text mining [188]. It is pre-trained on large-scale biomedical corpora, such as PubMed abstracts and full-text articles, making it particularly effective for tasks in the biomedical domain.

Key Features of BioBERT

- **Biomedical Corpus**: BioBERT is pre-trained on a large corpus of biomedical texts, which allows it to capture domain-specific knowledge and terminology.
- **Fine-tuning for Biomedical Tasks**: BioBERT can be fine-tuned for various biomedical NLP tasks, such as named entity recognition, relation extraction, and question answering in the biomedical domain.
- **Improved Performance**: Due to its specialized training, BioBERT often outperforms general-purpose models like BERT on biomedical tasks. It can better understand the context and nuances of biomedical text.
- **Applications**: BioBERT has been used in a variety of applications, including drug discovery, clinical decision support, and biomedical literature mining. Its ability to understand complex biomedical text makes it a valuable tool in these areas.

In one of our proposed methods (ReOnto in Chapter 4, we utilized BioBERT to obtain the similarity score between path and relation names

3.3 Neuro-Symbolic Artificial Intelligence

Neuro-Symbolic Artificial Intelligence (NeSy AI) is a branch of AI that merges neural and symbolic approaches to maximize their benefits [189]. Here, "neural" involves artificial neural networks or connectionist systems. In contrast, "symbolic" refers to methods using explicit symbol manipulation. NeSy AI's potential lies in combining the strengths of both approaches. Neural methods offer training from raw data and fault tolerance. Symbolic methods provide high explainability, provable correctness, and easy integration of human expert knowledge.

By using symbolic techniques alongside machine learning, especially deep learning, NeSy AI aims to enhance the handling of unknown terms, training with small datasets, error recovery,

and system explainability, outperforming deep learning-only systems.

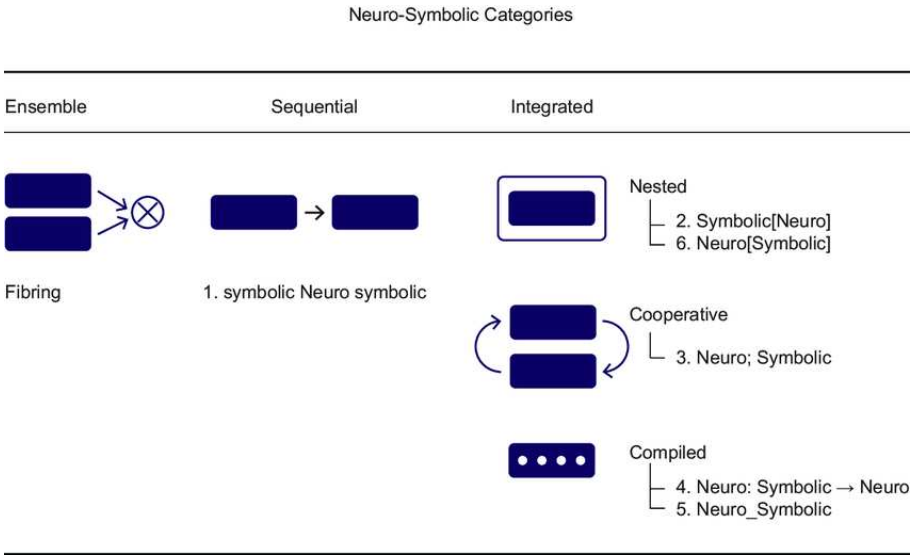

Figure 3.4: Neuro symbolic integration [3]

3.3.1 Classification of Neuro-Symbolic Artificial Intelligence

Henry Kautz, in his AAAI 2020 Robert S. Engelmore Memorial Award Lecture, discussed five categories of Neuro-Symbolic AI systems, depicted in Figure 3.4.

[symbolic Neuro symbolic] refers to an approach where input and output are presented in symbolic form, however all the actual processing is neural. This, in his words, is the "standard operating procedure" whenever inputs and outputs are symbolic.

[Symbolic[Neuro]] refers to a neural pattern recognition subroutine within a symbolic problem solver, with examples such as AlphaGo, AlphaZero, and current approaches to self-driving cars.

[Neuro $\cup$ compile(Symbolic)] refers to an approach where symbolic rules are "compiled" away during training, similar to the work on deep learning for symbolic mathematics [190].

[Neuro $\rightarrow$ Symbolic] refers to a cascading from a neural system into a symbolic reasoner, such as in the NeuroSymbolic Concept-Learner [191].

[Neuro[Symbolic]] refers to the embedding of symbolic reasoning inside a neural engine, where symbolic reasoning is understood as "deliberative, type 2 reasoning" as common, for example, in business AI, and including an internal model of the system's state of attention. Concepts are decoded into symbolic entities in an attention schema where attention to these concepts is high. A goal in the attention schema then signals that deliberative reasoning be executed.

CHAPTER 4

Knowledge enabled sentence-level relation extraction

Using existing RE techniques on biomedical text often produces unsatisfactory results because inferring relations directly from sentences is challenging due to the nature of biomedical relations. To address these issues, we present a novel technique called ReOnto, that makes use of neuro-symbolic knowledge for the RE task. ReOnto employs a graph neural network to acquire the sentence representation and leverages publicly accessible ontologies as prior knowledge to identify the sentential relation between two entities. The approach involves extracting the relation path between the two entities from the ontology. We evaluate the effect of using symbolic knowledge from ontologies with graph neural networks. Experimental results on two public biomedical datasets, BioRel and ADE, show that our method performs better than the baselines.

4.1 Introduction

In recent times, due to the exponential increase in data, knowledge bases have gained popularity as a means to efficiently store and organize information [192]. Although considerable efforts are invested in updating and maintaining knowledge bases, their incompleteness persists due to the dynamic nature of facts, which constantly evolve over the Web and other sources. Hence, there is a need to automate the process of extracting knowledge from text. In domains such as biomedicine, relation extraction task poses a few critical domain-specific challenges. Consider a sentence, *atrio ventricular (C0018827) conduction defects and arrhythmias by selective perfusion of a-v conduction system in the canine heart (C0018787)*, with entities C0018827 (ventricular) and C0018787 (heart) linked to UMLS [193]. Here target relation from BioRel dataset is *hasPhysicalPartOfAnatomicStructure*. The RE task aims to infer the semantic relationships. As demonstrated in the example, working with biomedical corpora poses several challenges. These include:

- Complex input sentences that may require extensive parsing and interpretation to extract relevant information.
- Indirectly inferred relations between entities in the text, which may require sophisticated natural language processing techniques.
- Difficulty obtaining domain knowledge of the specific entities mentioned in the text, which may require specialized expertise and additional research.

Moreover, in the biomedical domain, entities are intricately interlinked, resulting in numerous densely linked entities with high degrees and multiple paths connecting them [194]. Hence, inferring the correct relation from a given sentence may require reasoning about the potential path.

Limitations of Existing Works and Hypothesis. The existing approaches employ various techniques for relation extraction, such as multi-task learning [17] and transformers [18]. Graph Neural Network (GNN) models have been used to process complex relationships between entities [19, 20]. However, these deep learning models, which can incorporate semantic information about entities, often need large amounts of labeled data and computational resources to perform well [21, 22].

Albeit effective, these models employ standard message-passing or attention-based approaches (transformers, GNNs) which are inherently focused on homophilic signals [23, 24] (i.e., only on neighbourhood interactions) and ignore long-range interactions that may be required to infer the semantic relationship between two biomedical entities. Furthermore, sufficient domain-specific knowledge is available in various biomedical ontologies to be used as background knowledge for relation extraction.

It is also evident in the literature that reasoning over ontologies [28, 29] allows for the capture of relationships between entities that are distant from each other within the data [30, 31], which further helps in making predictions. For instance, in [32], ontology information was utilized as a tuple and transformed into a 3-D vector for predicting compound relations.

Hence, it remains an open **research question**: for biomedical relation extraction, can we combine reasoning ability over publicly available biomedical ontologies?

Contributions: To tackle this research question, our approach represents the first neuro-symbolic method for extracting relations in the biomedical domain. According to the classification of neuro-symbolic integration by Henry Kautz and Hamilton, our approach falls under the [Neuro[Symbolic]] category because our approach contains both neural networks and symbolic reasoning. Our method is two-fold. Firstly, we aim to aggregate the symbolic knowledge in the form of axioms (facts) consisting of logical constructs and quantifiers such as *there exist*, *for all*, *union* and *intersection* between entities present in various public ontologies and build background knowledge. In the second step, we incorporate background knowledge into a Graph Neural Network (GNN) to enhance its capabilities to capture long-range dependencies. The rationale behind using a GNN is to exploit the correlations between entities and predicates due to its message-passing ability between the nodes. Inducing external symbolic knowledge makes our approach transparent as we can backtrack the paths used for inducing long-range dependen-

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

cies between entities. Hence, we empower the GNN by externally induced symbolic knowledge to capture long-range interactions needed to infer biomedical relations between two given entities and a sentence. We name our approach as "ReOnto" with the following key contributions.

- Our novel relation extraction method, ReOnto, utilizes an ontology model to learn subgraphs containing expressive axioms connecting the given entities. It consists of a symbolic module incorporating domain-specific knowledge into a GNN, enabling the prediction of required relations between two entities within a biomedical knowledge graph.
- We study the effect of symbolic knowledge on the performance of the underlying deep learning model by considering several key characteristics such as 1) entity coverage from ontology, 2) the number of hops, etc. We provide conclusive evidence that aggregating knowledge from various sources to build the symbolic component (instead of using just one ontology for background knowledge) has a positive impact on the overall performance.
- We provide an exhaustive evaluation on two standard datasets, and our proposed method performs better than the baselines for biomedical relation extraction.

4.2 Problem Formulation and Approach

We define a Graph as a tuple $G = (\mathcal{E}, \mathcal{R}, \mathcal{T}^+)$ where $\mathcal{E}$ denotes the set of entities (vertices), $\mathcal{R}$ is the set of edge labels, and $\mathcal{T}^+ \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is a set of all labelled edges. The *RE Task* aims to find the target relation $r^c \in \mathcal{R}$ for a given pair of entities $\langle e_i, e_j \rangle$ within the sentence $\mathcal{W}$. If no relation is inferred, it returns *NA* label. In this section, we first discuss the ReOnto framework, which integrates graph neural network (GNN) [19] with symbolic knowledge. A GNN primarily employs three modules, which are encoding, propagation, and classification. Symbolic knowledge is integrated with the GNN score in the aggregation module (Figure 4.1).

4.2.1 Symbolic Module

As a first step, we aggregate symbolic knowledge (SK), available in public ontologies for extracting long-range dependencies between entity pairs. We build a connected graph G of the symbolic knowledge derived from ontologies. We define a graph $G = (V, E, T_+)$ where:

- V is the set of entities (vertices).
- E is the set of edges, where each edge $(v_s, r, v_o) \in E$ corresponds to a sequence $s = s_0^{s,o} s_1^{s,o} s_2^{s,o} \dots s_{l-1}^{s,o}$ extracted from the text. Here, v_s and v_o are the source and destination entities, respectively, and r is the edge label representing the relation between v_s and v_o.
- S_i represents a specific sequence or subsequence extracted from the text, associated with a pair of entities (v_s, v_o). The notation $s = s_0^{s,o} s_1^{s,o} s_2^{s,o} \dots s_{l-1}^{s,o}$ suggests that s is a sequence of elements (possibly tokens or words) related to the source entity s and the destination

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

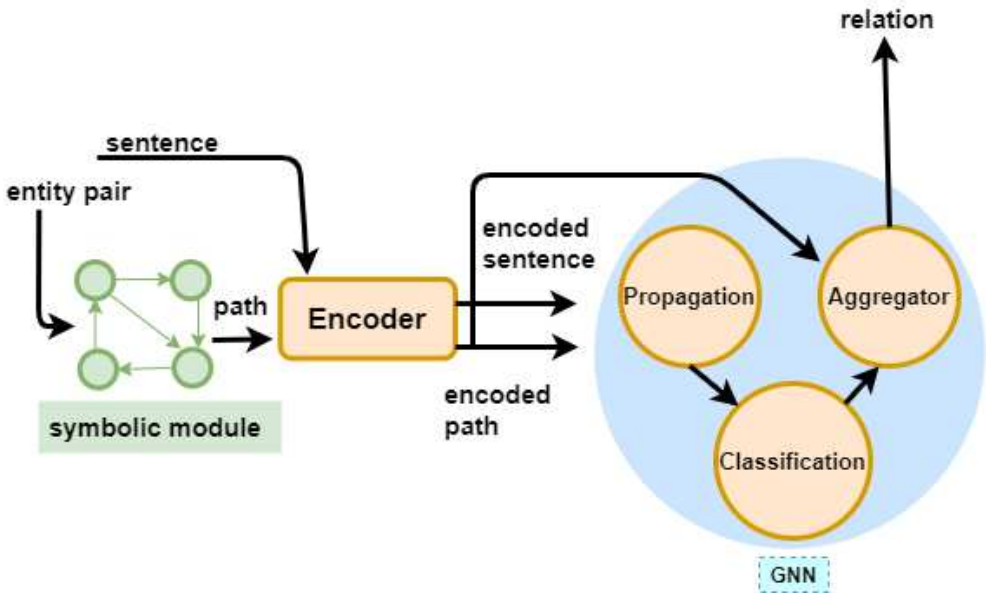

Figure 4.1: ReOnto Approach. The role of the symbolic module is to aggregate symbolic knowledge. It takes the entity pair and gives path information. 1) Encoding module accepts input vectors of sentence and path information to provide transition matrix. 2) Propagation module shares the hidden states of generated transition matrix with its neighbours 3) Classification module provides scores of prediction 4) Aggregator module integrates the score of the biased relation (from ontology reasoning) with that of the one from GNN to calculate loss.

entity o.

We also consider related SK of entity pair $SK^{s,o}$ which consist of path information ($\sum path_0^i$;$\sum axiom\_path_0^i$) $\in SK^{s,o}$ where i is the number of hops traversed to get the path. Path consists of multi hop details, each containing detailed information, while the axiom path contains path information enriched with expressive axioms. We identify, directly and indirectly, connecting paths between the entities v_s and v_o (Algorithm 2).

Single hop. For retrieving the direct path, we query on the ontology using SPARQL to check if a path exists between entity pair (e_s, e_o). The study examined the potential interactions in a sentence *Sandimmun, a medication formulated as cyclosporin (cya) in cremophor and ethanol, and the muscle relaxants atracurium and vecuronium in anesthetized cats.* The correct relation label between sandimmun and cyclosporin is *hasTradename*. Upon querying this entity pair from the ontology, it was found that the direct path between given entity pairs is *synonymOf* relation which is similar to the correct relation label *hasTradename* present in the dataset. The details on how we calculate the similarity is given is sec This highlights the similarity between *hasTradename* and *synonymOf*, as both relations indicate a form of equivalence or alternative naming. Details on how we calculate similarity is given in section 4.2.3. As depicted in Figure 4.2, the direct path between two given entities (if they exist) is extricated using

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

path$(y; e) \rightarrow$ cui$(x; y) \sqcap$ edge$(x; z) \sqcap$ cui$(z; e)$, where cui is concept unique identifier which uniquely identifies entity (assuming x is entity1, y is cui of entity1, z is entity2 and e is cui of entity2). It retrieves the connecting edge between two given entities. Once assimilating the path *synonymOf* between entity pairs, the aggregator module in ReOnto computes the similarity between the extracted path and all relations, assigning the correct label *hasTradename* as the similarity score reaches its maximum. The details of how to compute the similarity can be found in subsection 4.2.3

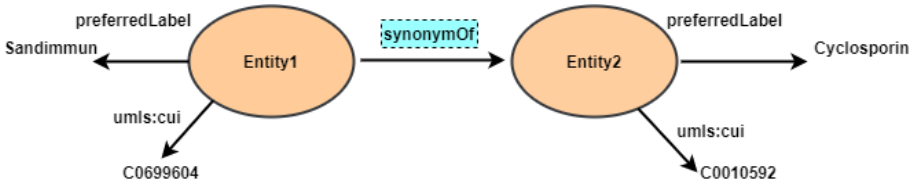

Figure 4.2: Subgraph of ontology illustrating direct connection between two entities

Multi hop. Multi-hop path reasoning over the knowledge base aims at finding a relation path for an entity pair by traversing along a path of triples from graph structure data [195]. For retrieving indirect path relation, we query on the ontology if a n-hop distance path exists between entity pair (e_s, e_o) starting from 1-hop distance path. Consider the sentence *Intravenous azithromycin-induced ototoxicity* with its relation label as *hasAdverseEffect*. From ontologies, we get the path as a concatenation of *causative agent of, has adverse reaction* using path(y;e)$\rightarrow$ cui(x;y) $\sqcap$ edge(x;z) $\sqcap$ edge(z;a) $\sqcap$ cui(a;e). The aggregator module receives this path as input and using a similarity score, assigns the target relation label *adverseEffect*. Refer to Figure 4.3 for details. Details about aggregate module is given in section 4.2.3

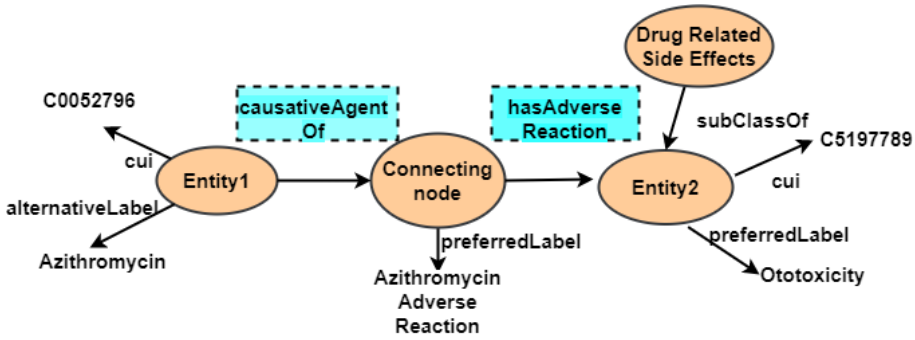

Figure 4.3: Subgraph of ontology depicting two hop distance between two entities

Using axioms. So far, we have considered only shallow and transitive relationships among the concepts. However, the biomedical domain consists of several complex relations. We argue that those relations can be captured using expressive axioms from ontology. Expressive axioms consist of logical quantifiers such as *there exist*($\exists$), *for all*($\forall$), *union*($\sqcup$), *intersection*($\sqcap$) which

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

are part of popular biomedical ontologies. These expressive axioms enrich an ontology and play an essential role in the performance of downstream applications [30]. Our objective is to determine the relation label between two entities by tracing the corresponding multi-hop triplet path that contains these axioms, starting from the first entity in the graph and continuing up to a specified distance until we reach the second entity. Note that when multiple paths are available between two entities, we have taken into account all the paths that are available which consist of unique keywords Consider the sentence, *A 73-year-old woman presented with fever and cough 2 weeks after completing the third cycle of fludarabine for chronic lymphocytic leukemia*. Here, correct relation label is *adverse effect*. From one of the ontology described in Table 4.2, we get following sub-graph enriched with axioms.

$$\textit{Fludaraline} \xrightarrow{\textit{causativeAgentOf}} \textit{Fludaraline Adverse Reaction}$$
$$\textit{Fludaraline Adverse Reaction} \sqsubseteq \exists \textit{hasFinding.Finding}$$
$$\textit{Cough} \sqsubseteq \textit{Finding}$$

From the above relations, one can see that Fludaraline and Fludaraline Adverse Reaction (FADR) has a relation *causativeAgentOf*. Moreover, there exists a *hasFinding* relation between FADR and Finding. Therefore, with ontology reasoning, we can interpret that Fludaraline has an axiom path consisting of *causativeAgentOf, hasSomeFinding*, which is closest to the relation label adverseEffect. Similarly, consider another sentence, *concentrations were significantly related to the degree of apocrine differentiation of the tumour and, in a subset of the cancers, capacity to release gcdfp-15 was positively correlated with incidence of progestogen and androgen receptors*. The labelled relation for this sentence is has *nichdParentOf* .

$$\textit{Tumor} \xrightarrow{\textit{qualifierBy}} \textit{Diagnostic Imaging}$$
$$\textit{Diagnostic Imaging} \xrightarrow{\textit{allowedQualifier}} \textit{Neoplasms Neoplasms} \sqsubseteq \exists \textit{parent.Post-Traumatic Cancer}$$
$$\textit{Post-Traumatic Cancer} \sqsubseteq \textit{Cancer}$$

For the above case, the derived path is *qualifierBy, allowedQualifier, subClass and there exist some parent and subClass*.

4.2.2 Encoding module

Entity pairs are encoded by concatenating the position embedding with the word embedding in the sentence (Equation 4.1), represented as $En(s_t^{s,o})$ where s_t is the word embedding and $p_t^{s,o}$ is the position embedding at word position t relative to the entity pair position (s, o). Similarly, symbolic path information from the Symbolic Module (SK) is encoded by concatenating path ($path_0^i$) and axiom path details ($axiom\_path_0^i$) where i represents the number of hops reaching destination.

$$En(s_t^{s,o}) = [s_t; p_t^{s,o}] \tag{4.1}$$

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

$$En(SK^{s,o}) = [\sum path_0^i; \sum axiom\_path_0^i]^{s,o} \tag{4.2}$$

The entity pairs representation and path information, after encoding with BioBERT are forwarded to a multi layer perceptron with non linear activation σ (Equation 4.3 and 4.4). We concatenate them as a shown in Equation 4.5. Since our dataset are from biomedical domain, we have used BioBERT [196] for encoding.

$$A_{s,o}^n = MLP_n(BioBERT(En(s_0^{s,o}), En(s_1^{s,o}), .., En(s_l - 1^{s,o})) \tag{4.3}$$

$$SP_{s,o}^n = MLP_n(BioBERT(En(SK^{s,o}))) \tag{4.4}$$

$$M_{s,o}^n = SP_{s,o}^{(n)} + A_{s,o}^{(n)} \tag{4.5}$$

4.2.3 Graph Neural Network

Propagation module

In this module, we propagate information among graph nodes where each node in the graph represents an entity using Equation 4.6, where given the representation of layer n , representation of layer $n + 1$ is calculated. Here n represents the index, B represents neighbours of v_o, and σ is the nonlinear activation function.

$$h_s^{n+1} = \sum_{v_o \in B(v_o)} \sigma(M_{s,o}^{(n)} h_o^{(n)}) \tag{4.6}$$

Classification module

In the classification module, embeddings of entity pair are the input. Now, ReOnto performs element wise multiplication on input and then passed into multi layer perceptron using Equation 4.7. Here $\cdot$ represent element wise multiplication.

$$MLP(v_s, v_o) = [h_{v_s}^{(1)} \cdot h_{v_o}^{(1)}]^T; [h_{v_s}^{(2)} \cdot h_{v_o}^{(2)}]^T; ...; [h_{v_s}^{(K)} \cdot h_{v_o}^{(K)}]^T \tag{4.7}$$

Aggregator module

Path information $(path_0^i; axiom\_path_0^i) \in SK^{s,o}$ from Symbolic Module is separately encoded using BioBERT [196] model, which is pre-trained on biomedical text corpora. At first, we perform the encoding of path information and total relation label R_1^i where i is the total number of potential relations (refer Equation 4.8 and 4.9). Then, we evaluate the semantic similarity score

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

using BioBERT model between path information and complete labelled relation list. We get the relation label with the maximum similarity score and add it as a weighted bias as given in Equation 4.10. An important observation to make is that the weights generated by the GNN undergo modification by incorporating the knowledge of the Symbolic Module. This step is crucial as it involves combining the symbolic and sub-symbolic components. We employ the softmax function to obtain probabilities and compute the cross entropy loss (refer Equations 4.11 and 4.12), where S denotes whole corpus and n are total entity pairs such that s $\neq$ o. It is worth noting that if no path exists between two entities, the bias score is set to 0, and loss is computed accordingly.

$$Renc = enc(R_1^i) \tag{4.8}$$
$$Penc = enc(SK^{s,o}) \tag{4.9}$$
$$biasedscore_r = max(cosSim(Renc, Penc)) \tag{4.10}$$
$$P(v_s, v_o) = softmax((MLP(v_s, v_o) + biasedscore_r) \tag{4.11}$$
$$L = \sum_{t=0}^{S} \sum_{s,o=0}^{n} (logP(v_s, v_o))_t \tag{4.12}$$

4.3 Experimental Setup

We conduct our evaluation in response to following research questions.

RQ1: What is the effectiveness of ReOnto that combines symbolic knowledge with a neural model in solving biomedical relation extraction task?

RQ2: How does knowledge encoded in different ontologies impact performance of ReOnto?

Datasets. Our initial biomedical dataset is BioRel [197], which includes a total of 533,560 sentences, 69,513 entities, and 125 relations types. The second dataset we use is the Adverse Drug Effect (ADE) dataset [198]. We approach the RE problem in this dataset as a binary classification task, where sentences are labeled as either positive (adverse-related) or negative (not adverse-related). Positive adverse relations are established when drug and reaction entities are associated in the given context, while negative relations involve drugs that are not accountable for a specific reaction. The ADE dataset comprises 6,821 labelled adverse sentences and 16,695 labelled negative adverse sentences, with a total of 5,063 entities. We consider two types of relations in this dataset: adverse-related and not adverse-related.

Table 4.1 provides details of the public ontologies utilized for constructing symbolic knowledge. DINTO (Drug to drug interaction ontology) designed to enhance the representation of drug-drug

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

interactions and their outcomes. OAE (ontology of adverse effects) focuses on representing and standardizing adverse events in medicine. NDF-RT (The National Drug File - Reference Terminology) provides a terminology for drugs and their characteristics. MEDLINE ontology is used for indexing and categorizing biomedical literature in the MEDLINE database. NCIt (National Cancer Institute Thesaurus) covers a wide range of cancer-related terminology and concepts.

Table 4.1: Ontologies used for Symbolic Knowledge

Ontology	Classes	Properties	Maximum depth
DINTO [28]	28,178	12	2
OAE [199]	10,589	123	17
NDF-RT [29]	36,202	90	9
MEDLINE [200]	2,254	12	2
NCIT [201]	177,762	97	21

Table 4.2: Hyper parameters setting

Hyper-parameters	Value
learning rate	0.001
batch size	50
dropout ratio	0.5
hidden state size	256
non linear activation	relu

4.4 Baseline Models for Comparison

We utilized several competitive baselines to evaluate the performance of our approach across different paradigms of relation extraction. These baselines are categorized as follows:

- **Multi-instance Models:** Multi-instance models are designed to handle scenarios where multiple sentences or instances are associated with a single relation label. These models aggregate information across instances to make predictions. We included widely-used multi-instance models such as CNN [202, 42, 203]. CNN-based models leverage convolutional layers to extract features from sentence-level representations, while PCNN introduces piecewise pooling to better capture localized information around entities. These models were adapted to the biomedical domain by re-training them on domain-specific datasets and incorporating biomedical context, such as entity descriptions and types, to ensure relevance to the task.

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

- **Sentential Relation Extraction (RE) Models:** Sentential RE models focus on extracting relations from individual sentences, often leveraging contextual information and graph-based representations. We included models such as Recon [24], GPGNN [20], and ContextAware [204]. For Recon [24], we used its entity attribute context (EAC) variant which encodes context from entity attributes ii) for fair comparison. GPGNN employs graph propagation mechanisms to model dependencies between entities, while ContextAware integrates surrounding context to improve relation prediction. These models were adapted to the biomedical domain by re-training them and inducing biomedical-specific features, such as entity descriptions and semantic types.
- **Biomedical Relation Extraction Works:** Biomedical RE models are explicitly designed for extracting relations in the biomedical domain, often leveraging domain-specific knowledge and embeddings. We included CRNN [205], CNN-Embedding [206], SparkNLP [207], RGCN [208], and BioRel [197]. CRNN combines convolutional and recurrent layers to capture both local and sequential dependencies, while CNN-Embedding integrates pre-trained embeddings tailored to biomedical text. SparkNLP utilizes advanced NLP pipelines for biomedical text mining, and RGCN employs relational graph convolutional networks to model complex entity relationships. BioRel incorporates domain-specific features and embeddings for enhanced performance. For these models, we used the reported values from their original papers, ensuring consistency with their published results.
- **T5 Model:** T5 [209] is a transformer-based model that has demonstrated state-of-the-art performance across various NLP tasks, including relation extraction. We fine-tuned T5 on biomedical datasets to adapt it to the domain, leveraging its ability to generate text-based outputs that capture relations effectively.

For models where code was publicly available, we executed them on both datasets to ensure a consistent evaluation framework. For biomedical-specific models, we relied on the reported values from their original papers to maintain fidelity to their published results. By adapting non-biomedical models to the domain and re-training them with biomedical context, we ensured a fair comparison across all baselines.

Hyper-parameters and Metrics. Table 4.2 outlines the best parameter setting. We employ GloVe embedding of dimension 50 for initialization. Since the datasets are from the biomedical domain for evaluating semantic similarity, we have used BioBERT model[1]. The size of position embedding is also kept at 50. We have used the open-source ontology (.owl) from BioPortal to extract the paths using the SPARQL query. We followed [20] for experiment settings. We evaluated our approach on both datasets using F1 score.

[1] `https://www.sbert.net/`

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

Table 4.3: Biomedical Relation Extraction Results. ReOnto outperforms baselines on both datasets. We've left the precision column blank for baselines that do not report it.

Dataset	Model	Accuracy (in%)	F1 scores
ADE	CNN [202]	68	0.71
	PCNN [203]	76.9	0.73
	ContextAware [204]	93	0.93
	RGCN [208]	86	0.83
	GPGNN [20]	92.1	0.90
	CRNN [205]	-	0.87
	CNN-Embedding [206]	-	0.89
	SparkNLP [207]	-	0.85
	T5 [209]	92	0.86
	RECON [19]	93.5	0.92
	ReOnto **(Ours)**	**97**	**0.96**
Dataset	Model	Accuracy (in%)	F1 scores
BioRel	CNN [202]	48	0.47
	PCNN [203]	64.6	0.57
	RGCN [208]	72	0.78
	GPGNN [20]	85	0.84
	CNN+ATT [197]	-	0.72
	PCNN+AVG [197]	-	0.76
	RNN+AVG [197]	-	0.74
	ContextAware [204]	89	0.87
	T5 [209]	88	0.86
	RECON [19]	89.6	0.86
	ReOnto **(Ours)**	**92**	**0.90**

4.5 Results

ReOnto performs better than the baseline models on both datasets (From Table 4.3). These results indicate that our model could successfully conduct reasoning with a neuro-symbolic graph on the fully connected graph and combine it with the underlying deep learning model (GNN in our case). Observed results successfully answer **RQ1**. Methods such as [19, 204] use contexts such as entity types and descriptions. Similarly, RECON and T5 include additional explicit information of long entity descriptions, its type that allows offline learning of entity context. However, in a real-world setting of the biomedical domain, it is viable that such context may

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

not be present for each entity. In contrast, our model discards the necessity of available entity context and learns purely using reasoning over connected entity graphs. Furthermore, multi-instance baselines try to learn relations using previous occurrences of entities in the document. In both cases, missing reasoning to capture long-range dependencies of entities hampers their performance. One possible reason why CNN and PCNN do not perform well is that biomedical sentences are complex, making it challenging to directly identify and adhere to the relationships within the text. We can also notice that the context-aware model is performing better than multi-instance on these datasets because entity contexts are helping up to an extent. Presently, we have added context information(symbolic knowledge) via ontology into the model. If enough context details are given our model can work on generalised datasets as well. Figure 4.4 presents plots a, b, c, d, which depict the training and validation F1 scores on both datasets, while plots e, f show the loss graph. Our observations indicate that ReOnto delivers consistent performance on these graphs within the considered timeframe (Table 4.5).

4.6 Ablation study

4.6.1 Effectiveness of number of ontologies

To better understand the contribution of each ontology on ReOnto's performance, we conducted an ablation study. Table 4.4 presents a summary of our findings, which indicate a decrease in performance when considering individual ontologies. This validates our approach of merging knowledge from multiple ontologies to create symbolic knowledge.

For the ADE dataset, we have a lesser entity coverage of 22% using DRON ontology. However, we found that the performance improves when we increase the entity coverage by incorporating the OAE and DINTO ontologies. This increase in entity coverage results in corresponding improvements in F1 scores. Similarly, for the BioRel dataset, we tested with MEDLINE ontology with entity coverage of 42% and then NCIt ontology with coverage of 34%, leading to corresponding improvements in F1 scores. Results also provide conclusive evidence that ReOnto's performance depends on the coverage of entities aligned with the dataset and combining encoded knowledge has positive impact on overall performance (answering **RQ2**).

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

Table 4.4: Effect of ontology on F1 scores

Dataset	Ontology	Entity coverage(approx.)	F1 scores
ADE	DRON [28]	22%	0.92
	OAE [199]	34%	0.93
	DINTO [210]	41%	0.95
BioRel	MEDLINE [200]	42%	0.88
	NCIt [201]	34%	0.84

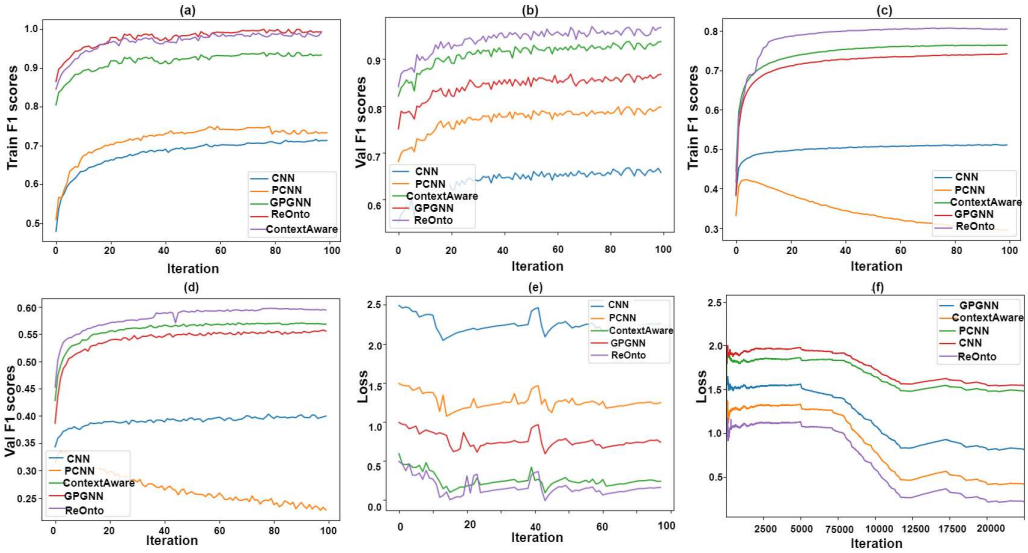

Figure 4.4: For the ADE dataset, Figures a) and b) show the training and validation F1 scores with baseline, respectively. Figure e) illustrates the cross-entropy loss for the iteration. For the BioRel dataset, Figures c) and d) show the training and validation F1 scores with baseline, respectively. Figure f) illustrates the cross-entropy loss concerning the iteration. ReOnto exhibits consistent and stable performance on both datasets, as indicated by the plotted F1 scores and loss.

4.6.2 Effectiveness of number of hops

We separately study the effect of the number of hops on the performance of ReOnto. Figure 4.5 shows the impact of the number of hops on the model. Increasing hops initially improve F1 scores until reaching a plateau. This is because additional hops don't provide new relevant information. Table 4.6 summarizes the extracted hops from the MEDLINE ontology, supporting our observation. Interestingly, increasing hops leads to redundant information that doesn't contribute to performance. To maintain context and meaningful connections, we preserved multi-hop information up to five hops in our experiment. Furthermore, Table 4.5 illustrates the

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

relationship between ontology size, parsing time, and the number of hops, indicating an increase in time as hops increase.

Table 4.5: Time taken to parse ontology and evaluate respective path. Parsing time increase w.r.t size of ontology

Ontology	Size (in KB)	Time taken (in seconds)					
		Parsing	Direct hop	One hop	Two hop	Axiom path1	Axiom path2
OAE	9286	6.75	0.11	2.73	7.47	2.19	5.92
NDFRT	69387	123.11	1.21	0.003	7.629	44.79	103.36
DINTO	1,10,865	137.4	1.6	3.8	8.54	5.67	11.32
MEDLINE	6975	2.19	0.002	0.0023	0.003	3.118	6.09
NCI	5,71,434	758.9	1034.5	1294.5	3454.1	2485	5569

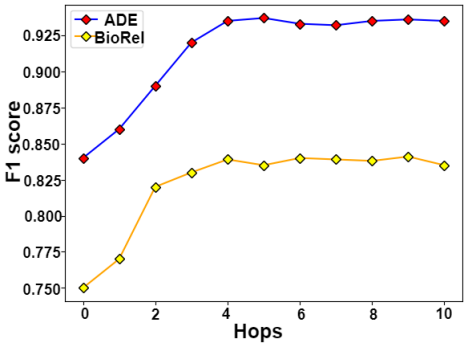

Figure 4.5: Effectiveness of hops on performance

Table 4.6: Derived path obtained by connecting “protein" and “dietary protein" entity

Hops	Path
path1	classifies
path2	mapped from dietary proteins, classifies
path3	classifies proteins, classifies dietary proteins, classifies
path4	classifies proteins, classifies dietary proteins, mapped from dietary proteins, classifies
path5	classifies proteins, classifies dietary proteins, related to carbs, related to dietary proteins, classifies
path6	classifies Proteins, classifies dietary proteins, mapped from dietary proteins, related to carbs, related to dietary proteins, classifies
path7	classifies Proteins, classifies dietary proteins, mapped from dietary proteins, related to carbs, related to dietary proteins classifies dietary proteins

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

Table 4.7: Sample sentences and predictions of various models. ReOnto using reasoning is able to predict the relations which are not explicitly observable from the sentence itself and requires long-range entity interactions.

Sentence	**Relation**	**GPGNN**	**Context Aware**	**ReOnto**
Both compounds are equally potent in the stimulation of intestinal calcium transport , bone(C0262950) calcium mobilization , in the elevation of serum phosphorus , and in the healing of rickets(C0035579) in the rat	is primary anatomic site of disease	may be associated disease of disease	may be finding of disease	is primary anatomic site of disease
The ventricular effective refractory period, as well as the vt cycle length(C0042514), increased after propranolol(C0033497) and was further prolonged after the addition of a type i agent	may be treated by	may diagnose	may treat	may be treated by
dsip and clip [acth(18-39)] immunoreactive(ir) neurons and fibers were examined in the human(C0086418) hypophysis and pituitary stalk using immunmohistofluorescence and peroxidase(C4522012) antiperoxidase methods	is organism source of gene product	nichd parent of	organism has gene	is organism source of gene product

4.7 Case study

Table 4.7 shows qualitative results that compare the ReOnto model with the baseline models. We report a few results showing ReOnto can predict the relationship with reasoning. ReOnto retrieved the relevant derived path from the ontology in the first case. ReOnto implicitly learns from the facts and captures the derived path to provide the correct relation label, even if it is not explicitly mentioned as *isPrimaryAnatomicSiteOfDisease*.

$$CUI{:}C0262950 \xrightarrow{preferredLabel} Bone$$
$$Bone \sqsubseteq \exists anatomicSiteOfDisease.Rickets$$
$$Rickets \xrightarrow{CUI} CUI{:}C0035579$$
$$Bone \xrightarrow{semanticType} Anatomic\ Structure$$

In the second case, ReOnto produces the following path by utilizing the expressive axiom of the ontology. ReOnto captures the long-range dependencies between entities and provides the correct relation label.

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

$CUI{:}C0042514 \xrightarrow{preferredLabel} Ventricular\ Tachycardia$
$Ventricular\ Tachycardia \equiv Techycardia$
$Techycardia \sqsubseteq \exists mayBeTreatedBy.Propranolol$
$Propranolol \xrightarrow{CUI} CUI{:}C0033497$

In the last case study, several paths are derived from the ontology for the *Human* entity. It can be observed that ReOnto derives and asserts the dependency path between *Human* and *Peroxidase*, and concludes that the target relation label *isOrganismSourceOfGeneProduct* applies, as compared to other baseline models. Such complex ontology reasoning provides long-range interactions between entities, which is inherently not possible in baseline models.

$CUI{:}C0086418 \xrightarrow{preferredLabel} Human$
$\exists Human \sqsubseteq geneProductHasOrganismSource.Myeloperoxidase$
$Myeloperoxidase \xrightarrow{hasDisposition} Peroxidase(disposition)$
$Peroxidase(disposition) \xrightarrow{preferredName} Peroxidase$
$Peroxidase \xrightarrow{CUI} CUI{:}C4522012$

4.8 Summary

In this chapter, we proposed a novel neuro-symbolic approach, ReOnto, that leverages path-based reasoning, including expressive axiom path with GNN. We apply our model to complex biomedical text and compare the approach with baselines. Based on the empirical results, there are three key takeaways. Firstly, existing baseline models with any form of context only capture short-range dependencies of entities. In contrast, our model uses long-range entity dependencies derived from ontology reasoning to outperform all baselines on both biomedical datasets. ReOnto provides effective reasoning on given text and entity pair, which can tackle the challenges of biomedical text. It also considers expressive axioms of ontology to reason on RE. The aggregation of these axioms outperformed the baselines. As a next step, we can consider using background knowledge on unsupervised data. An ontology reasoner can be used to infer more paths, and perhaps these additional axioms can improve the performance further. The code used in this research is available under the Apache License, Version 2.0, at `https://github.com/kracr/reonto-relation-extraction`.

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

4.9 Limitation

While ReOnto demonstrates promising results in sentence-level relation extraction by leveraging neuro-symbolic knowledge, it is not without limitations. One key limitation of the approach is its dependency on the availability of entities within the ontology. The method relies on extracting relation paths between entities from publicly accessible ontologies, which means that the effectiveness of ReOnto is constrained by the comprehensiveness and coverage of the ontologies used.

If an entity or its corresponding relation is not present in the ontology, the model may fail to infer the relation or provide suboptimal results. This limitation restricts the applicability of ReOnto to scenarios where entities and relations are well-represented in the ontology, potentially excluding novel entities or relations that are not yet documented. Additionally, the reliance on predefined symbolic knowledge may limit the model's ability to generalize to unseen or emerging biomedical concepts that are not captured in existing ontologies.

To address this limitation, future work could explore incorporating knowledge from diverse sources such as DBpedia, Wikidata, or other domain-specific Knowledge Graphs. By integrating multiple knowledge bases, the model could potentially expand its coverage and improve its ability to handle entities and relations that are not present in a single ontology. Furthermore, dynamically updating ontologies or employing techniques to infer relations for novel entities could enhance the robustness and generalizability of the approach.

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

Algorithm 1: Path generation via ontology

Input : entity pair (v_s,v_o), Number of hops (N)
Output: finalpath

Initialization:
    $i = 1$, source= v_s, $path_{i-1}$, $axiom\_path_{i-1}$, finalpath, adjacent node, hop_path_i, $axiom\_path_i$ = { }
finalpath = PathGeneration(v_s, v_o, N)
Function `PathGeneration(`v_s, v_o, N`)`**:**
    path, axiom_path, finalpath = { }
    foreach *entity pair* $v_s, v_o \in$ *ontology* **do**
        path.append(ExplorePath(v_s, v_o, N))
        axiom_path.append(ExploreSymbolicPath(v_s, v_o, N))
    finalpath = path $\cup$ axiom_path
    return {finalpath}
Function `ExplorePath(`v_s, v_o, N`)`**:**
    hop_path_i, adjacent node= GetNHopFromSource(v_s, 1) // calculates 1 hop distance from source
    $path_i$ = $hop\_path_i \cup path_{i-1}$
    if $v_o \neq$ *adjacent node and* $i \neq N$ **then**
        $path_i$= ExplorePath(adjacent node, v_o, N)
        $i = i + 1$
    return {$path_i$}
Function `ExploreSymbolicPath(`v_s, v_o, N`)`**:**
    $axiom\_path_i$, adjacent node = GetNHopFromSource(v_s, 2) // calculates 2 hop distance from source containing there exist and for all quantifier
    $axiom\_path_i$ = $hop\_path_i \cup axiom\_path_{i-1}$
    if $v_o \neq$ *adjacent node and* $i \neq N$ **then**
        $axiom\_path_i$= ExploreSymbolicPath(adjacent node, v_o, N)
        $i = i + 1$
    return {$axiom\_path_i$}

The work discussed in this chapter is part of the following publication. Monika Jain, Kuldeep Singh, and Raghava Mutharaju. 2023. ReOnto: A Neuro-Symbolic Approach for Biomedical Relation Extraction. In Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023.

CHAPTER 5

Knowledge enabled document-level relation extraction

Document-level relation extraction (DocRE) poses the challenge of identifying relationships between entities within a document as opposed to the traditional RE setting where a single sentence is the input. Existing approaches rely on logical reasoning or contextual cues from entities. We reframe document-level RE as link prediction over a knowledge graph with distinct benefits: 1) Our approach combines entity context with document-derived logical reasoning, enhancing link prediction quality. 2) Predicted links between entities offer interpretability, elucidating employed reasoning. We evaluate our approach on three benchmark datasets: DocRED, ReDocRED, and DWIE. The results indicate that our proposed method outperforms the state-of-the-art models and suggests that incorporating context-based link prediction techniques can enhance the performance of document-level relation extraction models.

5.1 Introduction

In recent years, document-level relation extraction problem (DocRE) evolved as a new subtopic due to the widespread use of relational knowledge in knowledge graphs [211] and the inherent manifestation of cross-sentence relations involving multi-hop reasoning. Consider the sentence in Figure 5.1 and its labelled relation *applies_to_jurisdiction* (Congress, US) from the DocRED dataset [83]. Even with the inclusion of multi-hop and co-reference reasoning, inferring the correct relation becomes challenging because the relation depends on multiple sentences and cannot be identified based on the language used in the sentence.

For these kinds of sentences, external context (knowledge) can play a vital role in helping the model capture more about the involved entities. For the above example, using the Wikidata knowledge base [212] and WordNet [213], we can get details, such as the entity types, synonyms, and other direct and indirect relations between entities (if they exist) on the Web As compared to traditional RE, DocRE has two major challenges: subject and object entities in a given triple might be dispersed across distinct sentences, and certain entities may have aliases in the form of distinct entity mentions. Consequently, the signal (hints) needed for DocRE is not confined to a single sentence. A common approach to solve this problem is by taking the input sentences and constructing a structured graph based on syntactic trees, co-references, or heuristics to represent relation information between all entity pairs [48, 42]. A graph neural network model is applied to the constructed graph, which performs multi-hop graph convolu-

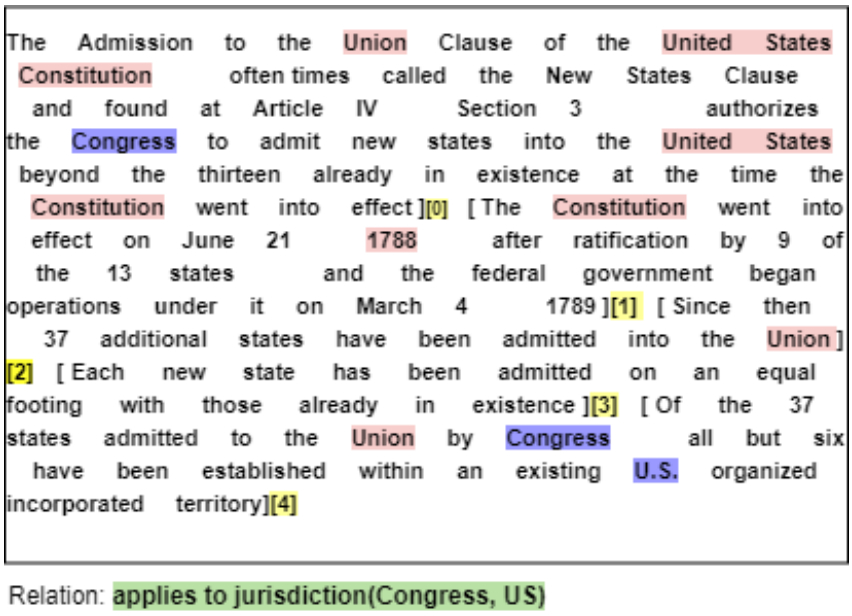

Figure 5.1: A partial document and labelled relation from DocRED. Blue colour represents concerned entities, pink colour represents other mentioned entities, and yellow colour denotes the sentence number.

tions to derive features of the involved entities. A classifier uses these features to make predictions [214]. Another approach [93, 215] explicitly models the reasoning process of different reasoning skills (e.g., multi-hop, coreference-mediated). However, even after considering features between the entity pairs and executing the reasoning process, DocRE is still hard due to the latent and unspecific contexts.

.

Previous research in this domain has underscored the potential of external context to enhance performance in relation extraction, co-reference resolution, and named entity recognition [216]. The distinctive innovation of our work lies in the fusion of context extracted from Wikidata and WordNet with a reasoning framework, enabling the prediction of entity relationships based on input document observations. Given the wide availability of external context, Knowledge Graph (KG) triples can augment training data, thereby casting the DocRE task as a knowledge Graph based link prediction challenge. In other words, given head and tail entities, we address the question of determining the appropriate relation.

We demonstrate that framing DocRE as a link prediction problem, combined with contextual knowledge and reasoning, yields enhanced accuracy in predicting the relation between entities. We furnish traversal paths as compelling justifications for relation predictions, thereby shedding

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

light on why a particular relation is favored over others. Notably, this marks the first instance of presenting a traversal path between entities for each prediction in the context of DocRE. According to the classification of neuro-symbolic integration by Henry Kautz and Hamilton, our approach falls under the [Neuro[Symbolic]] category since it involves both neural networks and symbolic reasoning. Our contributions in this work are as follows.

- We introduce an innovative approach named DocRE-CLiP (**Doc**ument-level

  Relation **E**xtraction with **C**ontext-guided **Li**nk **P**rediction), which amalgamates external entity context with reasoning via a link prediction algorithm.
- Our empirical analyses encompass three widely-used public document-level relation extraction datasets, showcasing our model's improvement over the recent state-of-the-art methods.
- For every prediction, our approach is first in DocRE literature to supply a traversal path as corroborative evidence, bolstering the model's interpretability.

5.2 Methodology

5.2.1 Problem Formulation

An unstructured document D consisting of K sentences is represented by $\{S\}_{i=1}^{K}$, where each sentence is a sequence of words. The entities within the document are denoted by $\mathcal{E} = \{e_i\}_{i=1}^{P}$ (where P is the total number of entities). Entities e_i has multiple mentions, $m_i^{s_k}$, scattered across the document D. Entity alias are represented as $\{m_i^{s_k}\}_{k=1}^{Q}$. Our objective is to extract the relation between two entities in $\mathcal{E}$ namely P(r|e$_i$,e$_j$) where $e_i, e_j \in \mathcal{E}, r \in \mathcal{R}$, here $\mathcal{R}$ is a total labelled relation set. The context (background knowledge) of an entity e_i is represented by C_{e_i} and a context path, i.e., a sequence of connected entities and edges from the head entity (e$_i$) to the tail entity (e$_j$) is represented by CP_{e_i,e_j}.

5.2.2 Approach

Our proposed framework, DocRE-CLiP, integrates document-derived reasoning with context knowledge using link prediction. In the first step, we extract triples from the sentences of the given document. In the second step, we extract two types of context: 1) entity context, such as its aliases, and 2) context paths from an external KB (Wikidata, in this case). Using the triples and extracted contexts, we create a context graph to calculate a link prediction score. Then in the third step, we use several reasoning mechanisms such as logical reasoning, intra-sentence reasoning, and co-reference reasoning to calculate relation scores for pairs of entities. In the

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

final step, the aggregation module combines the relation scores from the second and third steps. We have also implemented a path-based beam search in the framework to explain the predicted relation by providing traversal paths based on scores (refer to Figure 5.3). We now detail the architecture of our proposed framework.

Triplet extraction module. Document Relation Extraction (DocRE) datasets often contain labelled triplets; however, popular datasets [83] have about 64.6% missing triples, yielding an incomplete graph [217]. To extract all the triples from the document, we utilize an open-source state-of-the-art method [123] for triplet generation. This module takes a document D as input and produces triples (s, p, o), where s is the head entity, p is the relation, and o is the tail entity. These extracted triples (T) follow the Equation below, where n is the total count of triples extracted from document D.

$$T(D) = \{s_i, p_i, o_i\}_{i=1}^{n} \tag{5.1}$$

Context module. Our goal involves extracting two types of contexts – entity context and the contextual path between entity pairs. For entity e_i, we generate entity context using entity type and synonyms. This information is derived using WordNet [213]. We incorporate the entity context C_{e_i} into the triples. Here, S denotes the total number of extracted synonyms. Refer to Equations 5.2 and 5.3 for additional details.

$$C_{e_i} = \{e_i, hasSynonym, Synonym_k\}_{k=1}^{S}, \tag{5.2}$$

$$C_{e_i} = \{e_i, hasEntityType, EntityType\} \tag{5.3}$$

Let us consider an entity "Four" as an example. When we utilize WordNet, we discover that synonyms for "Four" include "4", "IV", and "quatern". Additionally, the entity is categorized as the type "number". Consequently, the following set of triples is generated through this process:

{Four, hasSynonym, 4}
{Four, hasSynonym, IV}
{Four, hasSynonym, quatern}
{Four, hasEntityType, number}

The second source of contextual information pertains to entity paths. Predicting relations between entity pairs poses challenges stemming from inherent document deficiencies [217, 115]. To address these issues, we introduced external context by harnessing insights from Wikidata. The procedure involves extracting paths (direct and indirect) between entity-entity, mention-

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

entity, and entity-mention pairs from Wikidata, provided they exist. The contextual path pertains to an entity pair e_i, e_j. We considered context paths spanning an N-hop distance (N being chosen based on experimental findings) between the entity pair. Subsequently, the extracted path is transformed into triples and forwarded to the link prediction model.

Illustrated in Figure 5.2 is the triple generation process employing a contextual path. The entity *Canadian* is two hops away from the entity *Ontario*. *Canada* is an intermediary entity, while *country* and *ethnic group* are intermediary properties. The Contextual Path (CP_{e_i,e_j}) are a set of triples formed using the intermediary entities and properties, as shown in Figure 5.2.

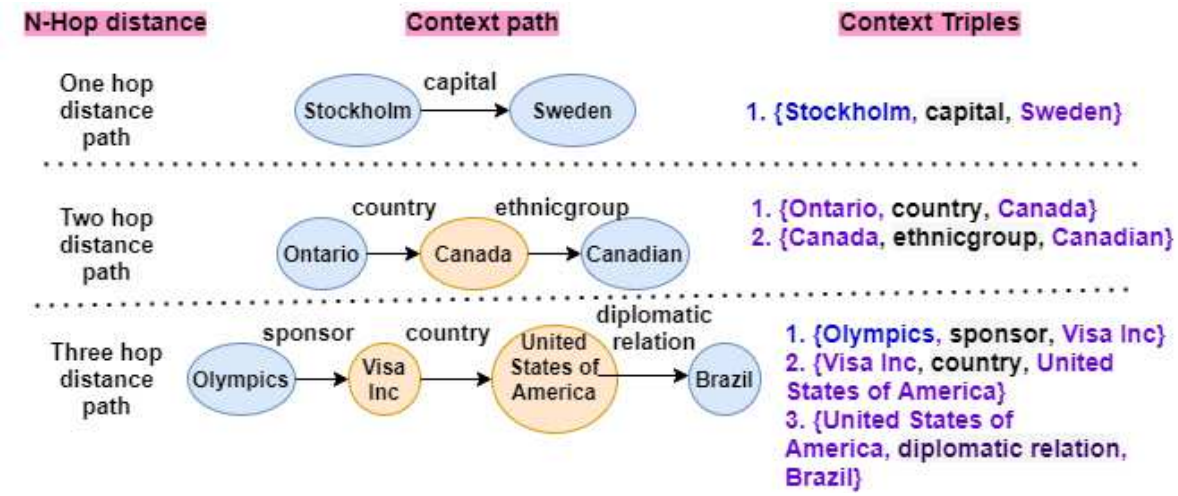

Figure 5.2: Triples constructed using N-hop path extracted from Wikidata. The head and tail entities are blue in colour. Intermediate entities are in peach colour.

Link prediction module. Link prediction is the task of predicting absent or potential connections among nodes within a network [218, 219, 220]. Given that the document relation extraction (DocRE) task involves constructing a graph that interlinks entities and considering our context, formulated as triples, which can be conceptualized as a Knowledge Graph (KG), we approach the DocRE challenge as a link prediction problem. This approach encompasses both an encoder and a decoder. The encoder maps each entity $e_i \in \mathcal{E}$ to a real-valued vector $v_i \in \mathbb{R}^d$, where $\mathbb{R}$ denotes the set of real numbers and d is the dimension of the vector. The decoder reconstructs graph edges by leveraging vertex representations, essentially scoring (subject, relation, object) triples using a function: $\mathbb{R}^d \times \mathcal{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$. While prior methods often employ a solitary real-valued vector $e_i \in \mathcal{E}$, our approach computes representations using an R-GCN [221] encoder, where $h_i^{(l)}$ is the hidden state of node e_i in the l-th layer of the neural network. To compute the forward pass for an entity e_i in a relational multi-graph, the propagation

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

model at layer $l+1$ is computed as follows.

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}(\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)})\right) \tag{5.4}$$

Here, $h_i^{(l)} \in \mathbb{R}^{d^{(l)}}$ with $d^{(l)}$ being the dimensionality of layer l. W_0^l and $W_r^{(l)}$ represent the block diagonal weight matrices of the neural network, and σ represents the activation function. N_i^r signifies the set of Neighbouring indices of node i under relation $r \in R$, and $c_{i,r}$ is a normalization constant.

For training the link prediction model, our dataset comprises a) core training triples from the dataset, b) triplets obtained through the triplet extraction module using Equation 5.1, c) triplets formulated using the context module guided by Equations 5.2 and 5.3, and d) triplets constructed using context paths connecting entity pairs. We use DistMult [222] as the decoder. It performs well on the standard link prediction benchmarks. Every relation r in a triple is scored using Equation 5.5.

$$P(r \mid i,j) = P(e_i^T \times R_r \times e_j) \tag{5.5}$$

Reasoning module. We consider three types of reasoning in our approach.
1) *Intra-sentence reasoning*, which is a combination of pattern recognition and common sense reasoning. Intra-sentence reasoning path is defined as $PI_{ij} = m_i^{s_1} \circ s_1 \circ m_j^{s_1}$ for entity pair $\{e_i, e_j\}$ inside the sentence s_1 in document D. $m_i^{s_1}$ and $m_j^{s_1}$ are mentions and "o" denotes reasoning step on reasoning path from e_i to e_j.
2) *Logical reasoning* is where a bridge entity indirectly establishes the relations between two entities. Logical reasoning path is formally denoted as $PL_{ij} = m_i^{s_1} \circ s_1 \circ m_l^{s_1} \circ m_l^{s_2} \circ\ s_2 \circ\ m_j^{s_2}$ for entity pair $\{e_i, e_j\}$ from sentence s_1 and s_2 is directly established by bridge entity e_l.
3) *Co-reference reasoning* which is nothing but co-reference resolution. Co-reference reasoning path is defined as $PC_{ij} = m_i^{s_1} \circ s_1 \circ s_2 \circ m_j^{s_2}$ between two entities e_i and e_j which occur in same sentence as other entity. Our implementation of these reasoning skills is inspired by [93].

Consider an entity pair $\{e_i, e_j\}$ and its intra sentence reasoning path (PI_{ij}), logical reasoning path (PL_{ij}) and co-reference reasoning path (PC_{ij}) in the sentence. The various reasoning is modeled to recognize the entity pair as intra-sentence reasoning $\text{R}_{PI}(r) = P\left(r \mid e_i, e_j, PI_{ij}, D\right)$, logical reasoning $\text{R}_{PL}(r) = P\left(r \mid e_i, e_j, PL_{ij}, D\right)$ and co-reference reasoning $\text{R}_{PC}(r) = P\left(r \mid e_i, e_j, PC_{ij}, D\right)$. Reasoning type is selected with max probability to recognize

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

the relation between each entity pair using the Equation:

$$P\left(r \mid e_i, e_j, D\right) = \max\left[\mathrm{R}_{PI}(r), \mathbf{R}_{PL}(r), \mathbf{R}_{PC}(r)\right] \tag{5.6}$$

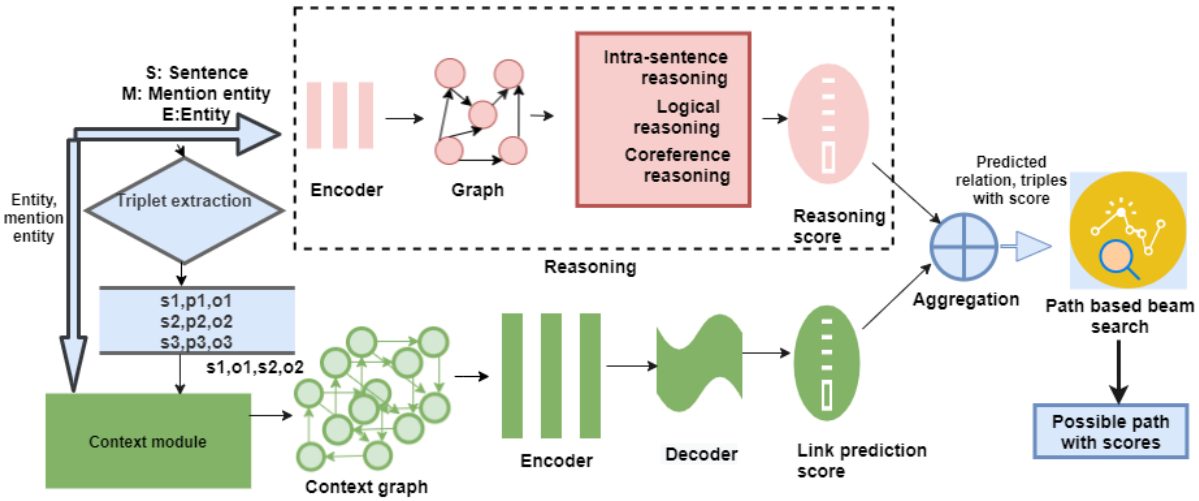

Figure 5.3: Illustration of proposed framework DocRE-CLiP and its various modules.

For discerning relations between two entities, we employ two categories of context representation – heterogeneous graph context representation (HGC) and document-level context representation (DLC) to model diverse reasoning paths [203, 95]. In heterogeneous graph context representation (HGC), a word is portrayed as a concatenation embedding of its word (W_e), entity type (W_t), and co-reference embedding (W_c). This composite embedding is then input into a BiLSTM to convert the document D into a vectorized form using the Equation: BiLSTM([W_e:W_t:W_c]). Following the methodology of [203], a heterogeneous graph is constructed based on sentence and mention nodes. For document-level context representation, following [223], a self-attention mechanism is employed to learn document-level context (DLC) for a specific mention based on the vectorized input document D.

To model intra-sentence reasoning path (α_{ij}), logical reasoning path (β_{ij}) and co-reference reasoning path (γ_{ij}), HGC and DLC representation are combined [224]. These reasoning representations are the input to the classifier to compute the probabilities of the relation between e_i and e_j entities by a multi-layer perceptron (MLP) for each path, respectively (Equation 5.7).

$$P\left(r \mid e_i, e_j\right) = \max\begin{bmatrix} \mathrm{sigmoid}\left(\mathrm{MLP}_r\left(\alpha_{ij}\right)\right), \\ \mathrm{sigmoid}\left(\mathrm{MLP}_r\left(\beta_{ij}\right)\right), \\ \mathrm{sigmoid}\left(\mathrm{MLP}_r\left(\gamma_{ij}\right)\right) \end{bmatrix} \tag{5.7}$$

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

By the end of this step, we will get the score of each relation (r) for a given $\{e_i, e_j\}$.

Aggregation module. In this module, we aggregate the probability score from the reasoning module Equation 5.7 and link prediction probability score using Equation 5.5. Further, the binary cross-entropy is used as a training objective function [83] for predicting the final relation.

5.2.3 Path-based beam search

An essential component of our approach is that it can explain the predicted relation by providing the most relevant path in the graph between the given entity pairs. This represents a notable advancement, as contemporary state-of-the-art models cannot often furnish explanations alongside their predictions. Unlike Greedy search, where each position is assessed in isolation, and the best choice is selected without considering preceding positions, we use beam search. This strategy selects the top "N" sequences thus far and factors in probabilities involving the concatenation of all previous paths and the path in the current position.

Inspired by [225, 226], we used beam search to derive plausible paths leading to the target entity within a graph. This graph (G) is constructed using the triples to train the link prediction module, augmented by test result triplets from the model's predictions. Our objective is to create a comprehensive graph encompassing the maximum available details, to generate substantial explanations for the predictions. Formally, we conceptualize the path-based beam search challenge as follows. Given a structured relational query (e_i, r, ?), where e_i serves as the head entity, r signifies the query relation, and (e_i, r, e_j) $\in$ G, our objective is to identify a collection of plausible answer entities e_j by navigating paths through the existing entities and relations within G, leading to tail entities. We compile a list of distinct entities reached during the final search step and assign the highest score attained among all paths leading to each entity. Subsequently, we present the top-ranked unique entities. This approach surpasses the direct output of entities ranked at the beam's apex, which often includes duplicates. Upon completing this step, we obtain actual paths (sequences of nodes and edges) for enhanced interpretability.

5.3 Experimental setup

We conduct our evaluation in response to the following research questions.

- **RQ1**: What is the effectiveness of DocRE-CLiP that combines context knowledge with reasoning in solving document-level relation extraction tasks?
- **RQ2**: How does knowledge encoded from external sources impact the performance of DocRE-CLiP?

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

- **RQ3**: Does the explanation generated by our approach provide sufficient grounds to support the inferred relation?

5.3.1 Datasets

The proposed model is evaluated on three widely-used public datasets 1) DocRED [83] 2) ReDocRED [96] and 3) DWIE [114]. ReDocRED is a revised version of handling DocRED issues such as false negatives and incompleteness [217]. Dataset details are in Table 5.1. For more information about these datasets refer chapter 2.

Dataset	#Triples	#Rel	#Entities	#Entity Types	#Doc
DocRED	50,503	96	30554	7	5053
ReDocRED	120,664	96	38239	10	5053
DWIE	19465	66	6644	10	777

Table 5.1: Dataset statistics

5.3.2 Baseline models for comparison

We used several competitive baselines and a recent state-of-the-art dataset for comparison. For the DocRED dataset, we compared our approach with various models, including both BERT-based and RoBERTa-based architectures.

BERT-based Models:

- **SIRE** [227]: This model focuses on improving the performance of document-level relation extraction by leveraging structured information and relational reasoning.
- **HeterGSAN-Rec** [224]: This approach utilizes a heterogeneous graph structure attention network to capture complex relationships within documents.
- **ATLOP** [95]: ATLOP employs an adaptive thresholding and localized context pooling mechanism to enhance the accuracy of relation extraction.
- **DRN** [93]: This model introduces discriminative reasoning techniques to better distinguish between different types of relations within documents.

RoBERTa-based Model:

- **DREEAM** [96]: DREEAM leverages the RoBERTa architecture to improve the robustness and generalization of relation extraction tasks.

Additional Comparisons:

- **KD-DocRE** [228]: This model incorporates knowledge distillation techniques to enhance the performance of document-level relation extraction.
- **DocuNet** [94]: DocuNet focuses on capturing long-range dependencies within documents to improve relation extraction.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

- **EIDER** [229]: EIDER uses an entity-aware inter-dependency reasoning approach to better understand the relationships between entities.
- **SAIS** [230]: SAIS employs a structure-aware information selection mechanism to improve the accuracy of relation extraction.

Similarly, for the ReDocRED dataset, we used the following models:

- **ATLOP** [95]: ATLOP employs an adaptive thresholding and localized context pooling mechanism to enhance the accuracy of relation extraction.
- **DRN** [93]: This model introduces discriminative reasoning techniques to better distinguish between different types of relations within documents
- **DocuNet** [94]: DocuNet focuses on capturing long-range dependencies within documents to improve relation extraction.
- **KD-DocRE** [228]: This model incorporates knowledge distillation techniques to enhance the performance of document-level relation extraction.
- **DREEAM**$_{inference}$ [96]: DREEAM leverages the RoBERTa architecture to improve the robustness and generalization of relation extraction tasks.

For the DWIE dataset, we considered the state-of-the-art models such as:

- **GAIN** [49]: This method leverages graph-based aggregation and inference mechanisms to extract document-level relations by reasoning over heterogeneous mention-level and entity-level graphs.
- **Joint embedding-BERT** [231]: This approach addresses the long-tail and multi-label challenges in document-level relation extraction by leveraging relation co-occurrence correlations through coarse- and fine-grained prediction tasks to enhance relation embeddings and improve relational fact extraction.
- **DRN** [93]: This model introduces discriminative reasoning techniques to better distinguish between different types of relations within documents

Additionally, we evaluated our model with context-based models such as:

- **KIRE** [100]: This framework enhances document-level relation extraction by injecting co-reference and factual knowledge from large-scale knowledge graphs, enabling improved reasoning and representation reconciliation for relational fact extraction.
- **RESIDE** [78]: A distantly-supervised relation extraction method that incorporates side information from knowledge bases, such as entity types and relation aliases, using Graph Convolution Networks to improve performance even with limited data.
- **RECON** [80]:A graph neural network-based method that aligns sentential relation extraction with knowledge graphs, leveraging entity attributes and factual triples to significantly improve extraction quality on benchmark datasets.
- **KB-graph** [99]: A joint information extraction model that integrates knowledge base entity representations via unsupervised entity linking, using attention-based and prior-based approaches to boost performance across named entity recognition, co-reference

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

resolution, and relation extraction tasks.

These models were compared with our proposed DocRE-CLiP approach to evaluate its effectiveness and competitiveness in document-level relation extraction tasks.

5.3.3 Hyper-parameters and Metrics

For the reasoning module, we follow the settings of [93]. We use the word embedding from GloVe (100d) and apply a Bidirectional LSTM (128d) to a word representation for encoding. We employ uncased BERT-Based model (768d) as an encoder with a learning rate 1e-3. We used AdamW as an optimizer, and the learning rate is $1e-3$. R-GCN is used as an encoder with a single encoding layer (200d) embeddings for the link prediction model. We regularize the encoder through edge dropout applied before normalization, with a dropout rate of 0.2 for self-loops and 0.4 for other edges. We apply l2 regularization to the decoder with a penalty of 0.01. Adam [232] is used as an optimizer, and the model is trained with 100 epochs using a learning rate of 0.01. For extracting the context paths, we use SPARQL queries to retrieve paths between entities. If multiple paths exist between entities, we consider the path with the highest page rank. The N-hop path length of the context varies from 1 to 4. The rationale behind this range is that we found no pertinent information for the context beyond four hops.

We use the evaluation metrics of DocRED [83], i.e., F1 and Ign F1 for DocRE-CLiP. Ign F1 is measured by removing relations in the annotated training set from the development and test sets.

5.4 Results

Table 5.2 compares DocRE-CLiP with various baseline models on DocRED, ReDocRED, and DWIE datasets. The results effectively address our primary research question (**RQ1**). To delve into the specifics of **(RQ1)**, we observe that incorporating context information from Wikidata and WordNet improves the performance compared to the baseline models. Notably, DocRE-CLiP surpasses all graph-based, reasoning-oriented, and transformer-based models by incorporating contextual information.

Examining the results for the DocRED dataset, our DocRE-CLiP model showcases an improvement of approximately 1% compared to top-performing models like KD-DocRE and DREEAM. For the ReDocRED dataset, DocRE-CLiP outperforms baseline models, including DREEAM, KD-DocRE, and DocuNet. Furthermore, in the case of DWIE dataset, our model outperforms all the baseline models, such as DRN, GAIN, and ATLOP, KIRE. Notably, the ReDocRED

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

dataset exhibits only slight improvement over the recent state-of-the-art DREEAM. This could be attributed to ReDocRED being an enhanced version of DocRED, which already tackled the issue of dataset incompleteness. On the other hand, both DocRED and DWIE demonstrate improvements by incorporating REBEL triplets (triples extracted using the REBEL model in the triplet extraction module), setting them apart from ReDocRED in this regard. The unique aspect of DocRE-CLiP lies in its incorporation of contextual paths between entities, which go beyond the entity context used in traditional context-aware approaches. Unlike methods such as KIRE, KB-both, and RESIDE, which primarily leverage knowledge bases to provide context, DocRE-CLiP utilizes a reasoning framework that explicitly models the relationships and interactions between entities through these contextual paths. This deeper integration of entity relationships contributes to its superior performance and sets it apart from other context-aware approaches.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

Table 5.2: Results on DocRED, ReDocRED, and DWIE datasets, including the baseline models. The precision column is blank (-) for baselines that do not report it. * denotes results obtained after modifying their code as the dataset necessitates. The mean and standard deviation of F1 and IgnF1 on the dev set are reported for three training runs. We report the official test score for DocRED on the best checkpoint on the dev set.

Dataset	Baseline	PLM/GNN	Dev		Test	
			F1	Ign F1	F1	Ign F1
DocRED	SIRE	BERT	59.82	61.6	60.18	62.05
	HeterGSAN-Rec	BERT	58.13	60.18	57.12	59.45
	ATLOP	BERT	59.22	61.09	59.31	61.30
	DRN	BERT	59.33	61.39	59.15	61.37
	DocuNet	RoBERTa$_{large}$	64.12	62.23	64.55	62.39
	ATLOP	RoBERTa$_{large}$	63.18	61.32	63.40	61.39
	KD-DocRE	RoBERTa$_{large}$	67.12	65.27	67.28	65.24
	SAIS	RoBERTa$_{large}$	65.17	62.23	65.11	63.44
	DREEAM	RoBERTa$_{large}$	67.41	65.52	67.53	65.47
	EIDER	RoBERTa$_{large}$	64.27	62.34	64.79	62.85
	KIRE	-	52.65	50.46	51.98	49.69
	RESIDE	GNN	51.59	49.64	50.71	48.62
	RECON	GNN	52.89	50.78	52.27	49.97
	KB-Graph	-	52.81	50.69	52.19	49.88
	DocRE-CLiP	BERT	**68.13**$_{\pm 0.15}$	**66.43**$_{\pm 0.17}$	**68.51**	**66.31**
DWIE	GAIN	BERT	62.55	58.63	67.57	62.37
	JE	BERT	63.38	58.40	69.12	62.92
	ATLOP	BERT	64.82	59.03	69.94	62.09
	DRN*$_{GloVe}$	BERT	-	-	56.04	54.22
	KIRE	-	65.62	56.58	67.37	58.41
	RESIDE	GNN	65.11	55.74	66.78	57.64
	RECON	GNN	65.48	56.12	66.94	58.02
	KB-Graph	-	65.39	56.03	66.89	57.94
	DocRE-CLiP	BERT	**66.12**$_{\pm 0.12}$	**57.11**$_{\pm 0.16}$	**67.10**$_{\pm 0.11}$	**58.87**$_{\pm 0.17}$
ReDocRED	ATLOP	BERT	-	-	77.56	76.82
	DRN*	BERT	-	-	75.6	74.3
	KD-DocRE	BERT	-	-	81.04	80.32
	DocuNet	RoBERTa$_{large}$	-	-	78.52	79.46
	DREEAM	RoBERTa$_{large}$	-	-	81.44	80.39
	DocRE-CLiP	BERT	-	-	**81.55**$_{\pm 0.14}$	**80.57**$_{\pm 0.22}$

5.5 Ablation study

Effectiveness of different contexts on DocRE-CLiP. Figure 5.4 offers an overview of our findings into DocRE-CLiP's performance under varying context conditions. Initially, we gauged

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

its performance without context and documented the outcomes. Subsequently, we introduced document triplets along with the dataset labels and determined the corresponding F1 scores. Additionally, we measured the impact of entity context and documented the ensuing performance. Furthermore, we scrutinized the effect of context path details on DocRE-CLiP by considering only those paths that notably enhanced performance. Our analysis establishes that the incorporation of context enhances performance across all datasets. Thus, we effectively address our second research question, **(RQ2)**. This study's findings lead us to conclude that DocRE-CLiP benefits the most from the context path compared to the other contexts.

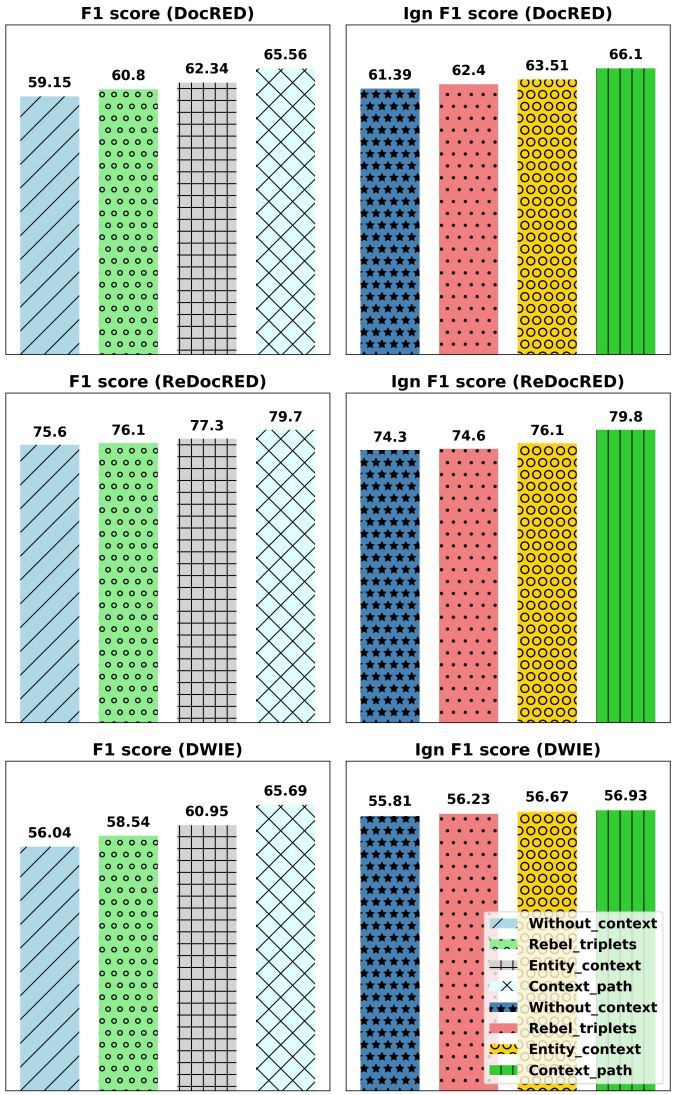

Figure 5.4: Performance of DocRE-CLiP across various contexts using the DocRED, ReDocRED, and DWIE datasets

Effectiveness of link prediction model with context. Our focus has been on investigating link prediction models utilizing individual triples from the dataset. Throughout our analysis,

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

we evaluate the performance of different link prediction models, specifically DistMult [222] , Complex [233], R-GCN [208], and KGE-HAKE [81]. Subsequently, we explored the influence of context on their performance. Notably, in each instance, upon the incorporation of context, the performance of the link prediction models improves. Table 5.3 summarizes the results. Considering these findings, RGCN has exhibited superior performance across metrics such as hits@1, hits@2, hits@10, and MRR. As a result, we have opted to select RGCN for link prediction.

Table 5.3: Performance of link prediction models.

Model	Metric	DocRED	ReDocRED	DWIE
DistMult	Hits@1	0.092	0.061	0.293
	Hits@3	0.104	0.088	0.307
	Hits@10	0.127	0.111	0.334
	MRR	0.105	0.080	0.307
DistMult+$_{context}$	Hits@1	0.10	0.071	0.297
	Hits@3	0.113	0.093	0.324
	Hits@10	0.112	0.12	0.337
	MRR	0.11	0.12	0.43
Complex	Hits@1	0.092	0.076	0.286
	Hits@3	0.097	0.096	0.296
	Hits@10	0.110	0.104	0.317
	MRR	0.099	0.087	0.297
Complex$_{context}$	Hits@1	0.101	0.09	0.31
	Hits@3	0.12	0.10	0.33
	Hits@10	0.15	0.13	0.36
	MRR	0.1	0.98	0.30
R-GCN	Hits@1	0.06	0.033	0.38
	Hits@3	0.09	0.06	0.43
	Hits@10	0.11	0.091	0.45
	MRR	0.0827	0.0532	0.40
R-GCN$_{context}$	Hits@1	0.11	0.51	0.67
	Hits@3	0.14	0.61	0.91
	Hits@10	0.23	0.13	0.98
	MRR	1.34	1.13	1.32
KGE-HAKE	Hits@1	0.07	0.05	0.44
	Hits@3	0.103	0.11	0.45
	Hits@10	0.123	0.112	0.47
	MRR	0.09	0.13	0.45
KGE-HAKE$_{context}$	Hits@1	0.10	0.08	0.48
	Hits@3	0.156	0.14	0.50
	Hits@10	0.18	0.144	0.51
	MRR	0.136	0.153	0.52

Study of the path for an explanation on DocRE-CLiP. Since generating explanations is a critical component of our model, we evaluate the effectiveness of the traversal paths used to support document-level relation extraction. Table 5.4 illustrates examples of explanations generated by our approach, showcasing how multi-hop reasoning enables the inference of relations. For instance, in the query *(IBM research Brazil, parent_organization, ?x)*, the explanation path

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

Query: (IBM research Brazil, parent_organization, ?x) **Answer**: IBM research **Explanation**: *IBM research Brazil, part_of, IBM research*
Query: (Piraeus, country, ?x) **Answer**: Greece **Explanation**: *{Piraeus, located_in_the_administrative_territorial_entity, Kiato}* *{Kiato, country, Greece}*
Query: (Quincy, country, ?x) **Answer**: United States **Explanation**: *{Quincy, country, American}* *{American, country_of_citizenship, America}* *{America, synonym, United States}*

Table 5.4: Example queries and results on DocRED dataset

identifies the intermediate relationship *(IBM research Brazil, part_of, IBM research)* to infer the correct answer, *IBM research*. Similarly, for the query *(Piraeus, country, ?x)*, the explanation path traverses through *(Piraeus, located_in_the_administrative_territorial_entity, Kiato)* and *(Kiato, country, Greece)* to infer the answer *Greece*. These examples demonstrate how the model leverages multi-hop reasoning to provide sufficient grounds for inferred relations, effectively addressing **RQ3**.

5.6 Case Study

We discuss two successful and one failed case of DocRE-CLiP and compare them with the baseline model, DRN (Table 5.5).

Case1: To identify the relation between Manche and France in sentence 0, we use external knowledge about France and Manche. We get connecting context path using DocRE-CLiP: *{France, contains_the_administrative_territorial_entity, Normandy}*
{Normandy, contains_the_administrative_territorial_entity, Mance}. Following the context path directly leads to the relation *contains_the_administrative_territorial* between France and Mance.

Case2: With the aid of context path between entities, *{United Kingdom, legislative_body, Parliament of the United Kingdom}* *{Parliament of the United Kingdom, instance_of, Parliament}*. DocRE-CLiP successfully identifies correct relation *legislative body*. **Case3**: Using pattern recognition, DocRE-CLiP identifies the publication date as a relation. However, the inclusion of the entity context, such as “Troublemaker”, “description”, and “song”, does not contribute significantly to the accuracy of the prediction. Consequently, DocRE-CLiP encounters difficul-

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

ties in correctly predicting the relation.

Case1:
Sentence:
Sentence 0 : The Château. de Pirou is a castle in the commune of Pirou in the département of Manche (Normandy) France]
Correct answer: contains_administrative_territorial entity
Baseline: country
DocRE-CLiP: contains_administrative_territorial_entity
Case2:
Sentence:
Sentence 0 : The Wigram Baronetcy of Walthamstow House in the County of Essex is a title in the Baronetage of the UNITED KINGDOM. Sentence 3 :The second Baronet also represented Wexford Borough in Parliament Sentence 5 : The fourth Baronet was a Lieutenant - General in the army and sat as a Conservative Member of Parliament for South Hampshire and Fareham
Correct answer: legislative_body
Baseline: has_part
DocRE-CLiP:legislative_body
Case3:
Sentence:
Sentence 1 :[Taking a more electronic music sound than his previous releases TY.O was released in December 2011 by Universal Island Records but for reasons unknown to Cruz its British and American release were held off] Sentence 3 :[TY.O features a range of top - twenty and top - thirty singles including “Hangover” (featuring Florida) “Troublemaker”“There She Goes” (sometimes featuring Pitbull) the limited release “World in Our Hands” and “Fast Car” which features on the Special Edition and Fast Hits versions of the album]
Correct answer: Inception
Baseline: publication_date
DocRE-CLiP: publication_date

Table 5.5: Case study with DocRE-CLiP prediction. Blue colour represents entities in the sentence, and purple colour represents the DocRE-CLiP prediction.

5.7 Additional experiments

Other than the evaluations discussed earlier in the chapter, we also evaluated our approach using the top hits from Wikidata. These top hits are calculated using Networkx[1] library, which follows the Google page rank. Using these triplets, we tested with link prediction models and evaluated the triple probability score. These triplets are scrutinized by setting a threshold of 0.9, and the F1 score is computed. Observing the hits of Wikidata, we found that most hits belong to the property called “alt-label”, which does not improve the F1 score. Therefore we omitted this property and re-evaluated the F1 score. Table 5.6 are the results obtained by using this entity context. After getting promising results, we considered the context path discussed in the chapter to improve the results further. We tried one-hop and two-hop distances individually and recorded the results.

[1] https://networkx.org/

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

Model	Top Hits (Wikidata)	Top Hits (Wikidata) without altLabel	One Hop	Two Hop
DistMult	0.31	0.330	0.341	0.371
complex	0.5	0.52	0.531	0.54
RGCN	0.328	0.34	0.410	0.421
KGE-HAKE	0.310	0.329	0.331	0.345

Table 5.6: Link prediction results using DocRED dataset

Furthermore, by varying the threshold from 0.5 to 1, we have observed that the F1 score corresponds to the threshold from the Complex model. The F1 score will be high when the threshold is 0.5, but as the threshold progresses to 1, the F1 score decreases. This suggests that lower thresholds capture a larger number of triples, which contributes to higher F1 scores, while higher thresholds filter out more triples, leading to a decline in the performance metrics. Figure 5.5 illustrates these results.

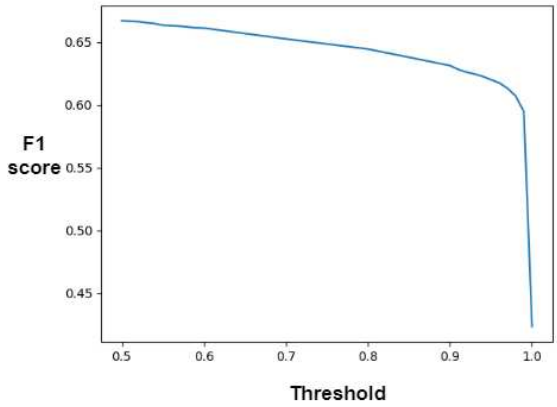

Figure 5.5: F1 score observed w.r.t threshold for complex model

5.8 Summary

This chapter introduces DocRE-CLiP, a context-driven approach for document-level relation extraction (DocRE). In this approach, the link prediction score is calculated using a context graph created from entity context and context paths derived from external knowledge. We assess our framework using the DocRED, ReDocRED, and DWIE datasets. DocRE-CLiP demonstrates superior performance across all datasets. Notably, DocRE-CLiP outperforms graph-based, reasoning-oriented, and transformer-based models. Our results suggest that integrating diverse context types into the link prediction module enriches relation prediction within the DocRE framework, providing interpretability. As future work, researchers can extend our efforts by crafting a versatile model capable of traversing diverse document types, thereby significantly

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

amplifying its aptitude for assimilating knowledge. Furthermore, with supporting evidence provided in our work as the first step, the document RE and KG link prediction research findings will mutually benefit.

The code used in this work is available under the Apache License, Version 2.0, at `https://github.com/kracr/document-level-relation-extraction`.

5.9 Limitation

Although the proposed method for document-level relation extraction (DocRE) shows strong performance and surpasses existing models, it does have certain limitations. One major drawback is that the model depends on both entities being present in the knowledge base (KB), specifically Wikidata, to extract their context and predict relationships. If one or both entities are missing from Wikidata, the model struggles to infer the relationship, which can lead to less accurate results. This reliance limits the model's effectiveness in cases where entities are either new, highly specialized, or not well-represented in Wikidata. Additionally, the approach may not work well in situations where only partial information about the entities is available. To overcome this limitation, future research could explore ways to use the context of one entity even if the other entity is missing from the KB. This could involve integrating information from other knowledge sources like DBpedia or domain-specific databases. Another potential direction could be developing techniques to predict relationships using incomplete or partial entity information, which would make the model more flexible and applicable to a wider range of scenarios.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru and Kuldeep Singh. 2024: Revisiting Document-Level Relation Extraction with Context-Guided Link Prediction. Main Technical Track. AAAI 2024. February 22-25, 2024.

CHAPTER 6

Knowledge enabled cross document-level relation extraction

Beyond a single sentence or a document, entity pairs can also be present in different documents [52]. The relationships between of entity pairs present in different documents is called cross document-level relation extraction. Current cross document-level relation extraction (CrossDocRE) efforts do not consider domain knowledge, which is often assumed to be known to the reader when documents are authored and read. Here, we propose a novel approach, KXDocRE, that embed domain knowledge of entities with input text for cross-document RE. Our proposed framework has three main benefits over baselines: 1) It incorporates domain knowledge of entities along with documents' text; 2) It offers interpretability by producing explanatory text for predicted relations between entities 3) It improves F1 score over the prior methods. We propose three variations of KXDocRE, i.e., entity type context (EC), connecting path context (CC) and entity type with connecting path context (ECC). In all three variations, it was able to perform better than baseline models.

6.1 Introduction

An analysis of Wikipedia reveals that over 57.6% of relational facts are not co-located within a single document [212]. This suggests that a significant portion of these facts is distributed across multiple documents. In essence, more than 57.6% of relationships between entities are not confined to a single document but are instead dispersed across various sources. This distribution highlights the fragmented nature of relational facts across different documents. Considering that, little work has been performed in cross-document relation extraction (CrossDocRE). CodRED (cross-document relation extraction dataset) is the first dataset published in this line of work [52], which serves as a starting point to solve CrossDocRE. Documents containing the source entity are identified and retrieved from multiple documents; the same is done for the target entity. Various text paths between entity pairs (source and target entity) across documents are recognized using the mentioned entities (other than the source and target entity). Text paths refer to the connections that link source and target entities through mentioned entities. These paths are retrieved from both the source and target documents. Knowledge graphs play a crucial role in providing the missing links, enabling the connection of these dots across different documents

Figure 6.1 shows an example of CrossDocRE. Between entity pairs **GCompris** and **GNU**

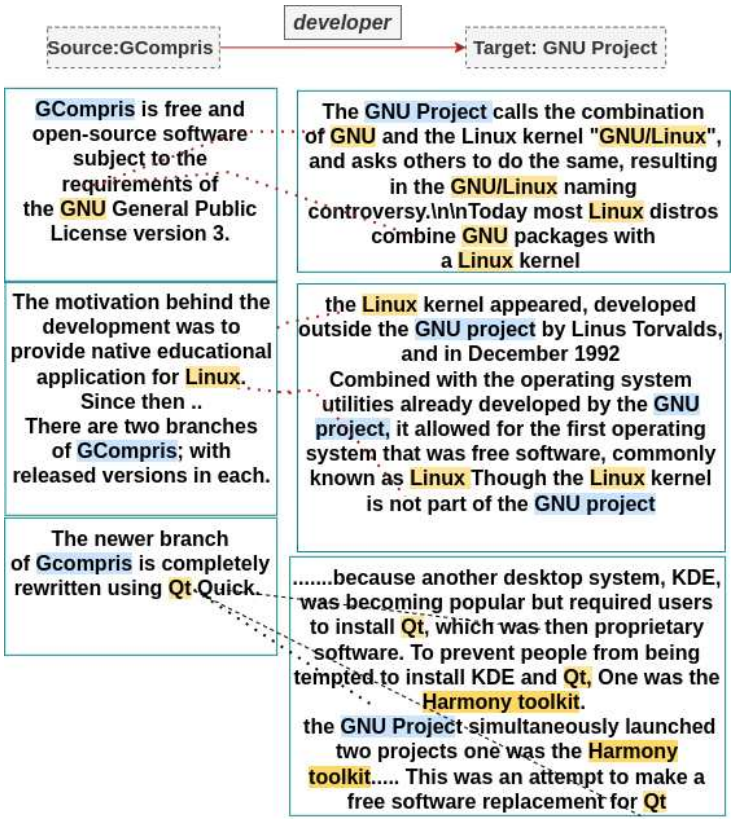

Figure 6.1: Three textual paths indicate the relationship path between the source entity (GCompris) and the target entity (GNU Project). These connections are established through pairs of documents, where one document features the source entity, and the other contains the target entity. In each path, the connection between the source and target entities is led by a mentioned entity in both documents (e.g., Linux).

Project, multiple text paths via the mentioned entities, such as *Linux, Qt* and *GNU*, can be used to get the correct relation label as **developer**. CrossDocRE identifies relationships between entities across multiple documents. It establishes textual paths using mentioned entities that appear in both documents. By leveraging information from multiple documents, CrossDocRE can capture complex relationships that might not be evident from a single document, thereby enriching the knowledge graph with more connections. CodRED uses these text paths in a bag and performs reasoning [52]. Although this approach seems reasonable, along with relevant text, these text paths also contain noisy data. To overcome these issues, ECRIM (entity-centered cross-document relation extraction) proposed filtering text paths using a mentioned entity [102]. However, ECRIM only works in one of the settings of CrossDocRE where text paths are provided for reasoning. To address this issues, MR.COD has been proposed, which is a multi-hop reasoning framework based on path mining and ranking [234]. These models rely on the knowledge between entities in a text and do not consider the domain knowledge associated with entities. Past work along these lines uses entity types and entity aliases to predict

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

the relation [79]. RECON [19] encoded attribute and relation triples in the Knowledge Graph. KB-Both [39] uses entity details from hyperlinked text documents of Wikipedia and Knowledge Graph (KG) from Wikidata to enhance performance. Given input text as a sentence or a document, these models use basic details of entities to improve the performance. In this chapter, we explore whether incorporating domain knowledge can enhance the performance of CrossDocRE tasks. According to the classification of neuro-symbolic integration by Henry Kautz and Hamilton, our approach falls under the [Neuro[Symbolic]] category since our approach contains both neural network and symbolic reasoning. The main contributions of this work are as follows.

- We propose a novel model that integrates domain knowledge with a cross-document relation extraction model.
- Our validation demonstrates the effectiveness of our model, with nontrivial performance gains in cross-document RE.
- We enhance the predicted relationships through textual explanations, offering insights into how the relations were predicted.

As mentioned in related work chapter 2, CodRED [52] is the first open source human-annotated cross-document dataset. In this work, the authors [52] address the problem using two approaches. The first method involves a pipeline approach in which they construct a relational graph for each document and reason over these graphs to extract the desired relation. The second method, referred as the joint approach, combines various text path representations through a selective attention mechanism to predict relations. Although this method is effective, it does not consider the mentioned entity-based sentences. ECRIM uses an entity-based document context filter to retain useful information in the given documents by using the mentioned entities in the text paths. Secondly, they solve CrossDocRE using cross-path entity relation attention, allowing entity relations across text paths to interact with each other [102]. Nevertheless, this work focuses on a closed setting where evidential context has been given instead of all documents. A multi-hop evidence retrieval method based on evidence path mining and ranking has been proposed in MR.COD [234]. In evidence path mining, a multi-document passage graph is constructed, where passages are linked by edges corresponding to shared entities. A graph traversal algorithm mines the passage paths from source-to-target entities. In evidence path ranking, paths are ranked based on relevance and top-K evidence paths are selected as input for downstream relation extraction models. Alternatively, a causality-guided global reasoning algorithm is also used to filter confusing information and achieve global reasoning to solve cross-document relation extraction [103]. Proposed models for CrossDocRE until now does not consider background knowledge.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

6.2 Methodology

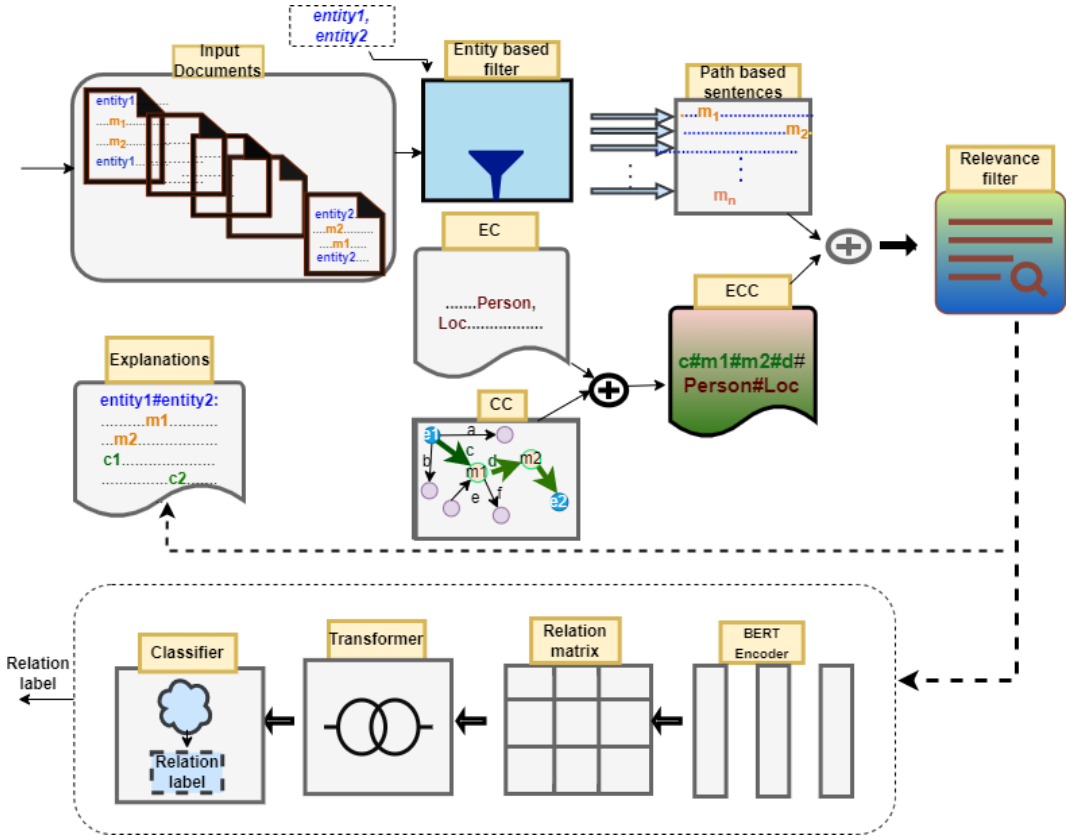

Figure 6.2: The architecture diagram of KXDocRE is designed for cross-document relation extraction. Initially, input documents containing source and target entities are processed. Sentences within these documents are filtered based on the presence of these entities. Additional content is incorporated using various settings such as EC (Entity Context), CC (Connecting Context), and ECC (Entity and Connecting Path Context). Following this, the filtered sentences are passed to the relevance filter module. This module assesses the semantic relevance of the sentences. The relevant sentences are then processed by a transformer, which ultimately outputs the relation label

.

6.2.1 Problem statement

For a given entity pair, $\langle e_s, e_o \rangle$ our task is to predict the target relation $r^c \in \mathcal{R}$ that holds between e_s and e_o within a given corpus of documents $\mathcal{C}_{i=1}^{n}$, where $\mathcal{R}$ is the relation set and n is total number of documents. If no relation is inferred, it returns the *NA* label. Besides e_s and e_o, we introduce the notion of potential bridge entities, which are other mentioned entities in a document that could act as intermediate concepts that help link e_s and e_o via multi-hop connections. The documents provided in CrossDocRE are also annotated with these mentioned entities. We denote $E_i = \{e_i^1, \ldots, e_i^m\}$ to be the set of m mentioned entities in the i-th docu-

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

ment, $i = 1, \ldots, M$. CrossDocRE works in two settings, closed and open. In the closed setting, only the related documents are provided to the models, and the relations are inferred from the provided documents. In the more challenging and realistic open setting, the whole corpus of documents is provided, and the model needs to efficiently and precisely retrieve related evidence from the corpus. We now discuss the architecture of KXDocRE (Figure 6.2) in the subsequent sections.

6.2.2 Domain Knowledge

We consider three forms of domain knowledge for cross-document relation extraction [235]: 1) The type information of the source and target entity; 2) The connecting path between the source and target entities, extracted using Wikidata; and 3) A combination of 1) and 2). We use the term *context* in the rest of the chapter to capture these three forms of knowledge.

Entity type context (EC): Predicting the relationship between two entities relies heavily on their respective entity types. For instance, if both entities belong to the *Person* category, certain relationships such as *has organization* and *has location* are not viable. However, relationships such as *child* and *spouse* become possible. Therefore, incorporating entity type information can assist the model in excluding obvious relationships, thereby improving its performance. Consider the example given in Figure 6.3. The type of the source entity *Jim Lynagh* is *Person* and the type of the target entity *Irish Republic*, is *Geo Political Entity*, which is used to identify locations, countries, cities, or geopolitical regions. The entity type context for this example is, $\text{EC}_{\{\text{Q6196505, Q1140152}\}}$={Person, GeoPoliticalEntity}

Connecting path context (CC): The contextual path pertains to an entity pair $\langle e_s, e_o \rangle$. We consider context paths of length up to N_h, the hop distance (a tunable parameter) between the entity pair. In Figure 6.3, the entity *Jim Lynagh* is five hops away from the entity *Irish Republic*. The four nodes between the entity pair on the path are intermediary entities, and the five edges are the intermediary properties. The contextual path (CP_{e_i,e_j}) is formed using the intermediary entities and properties. So, the connecting path context, $\text{CC}_{\{\text{Q6196505,Q1140152}\}}$={instance of, Human, model item, Douglas Adams, country of citizenship, United Kingdom, replaces, United Kingdom of Great Britain and Ireland, followed by}.

Entity type with connecting path context (ECC): ECC combines the entity type and the connecting path context. For Figure 6.3, $\text{ECC}_{\{\text{Q6196505,Q1140152}\}} = \text{EC}_{\{\text{Q6196505,Q1140152}\}} \cup \text{CC}_{\{\text{Q6196505,Q1140152}\}}$.

The steps to generate EC, CC, and ECC for a given entity pair are outlined in Algorithm 2. The ContextGeneration function (lines 3-12) computes the EC, CC, and ECC for the given entity pair. CC is constructed in steps of one hop. In lines 14-17, the adjacent edge and node

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

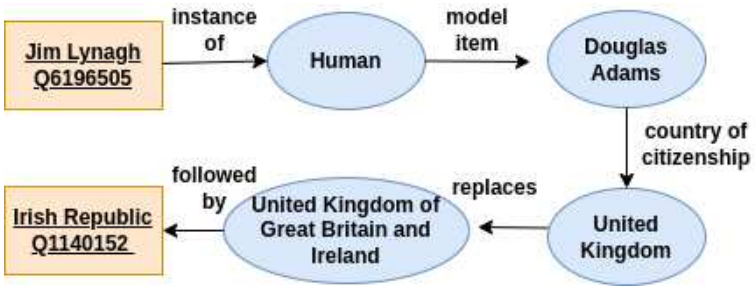

Figure 6.3: Context path constructed from Wikidata between Jim Lynagh (source) and Irish Republic (target).

information is retrieved and added to the path. This continues for N_h hops (lines 18-20). The EC of the source and the target entity is obtained using the named entity recognition technique (lines 27-35).

6.2.3 Entity-based filter

In this step, we select sentences that contain meaningful connecting information about the source and target entity and remove irrelevant sentences. The sentences are filtered using the mentioned entity, e^m. Similar to the baseline [102], our underlying premise is that a sentence holds significance if it contains either a source or a target entity. Additionally a sentence is considered more significant if, in addition to mentioning the pair of entities present in both documents, it also refers to another related entity. We calculate each entity's score based on three conditions. 1) The source and target entities co-occur in some sentence with e^m (Θ_1), 2) An entity, e^o, co-occurring with e^m, also co-occurs with the source entity in a sentence and the target entity in a different sentence (Θ_2), 3) e^m is part of a text path connecting source and target entities (Θ_3). Figure 6.4 depicts these conditions using the example from Figure 6.1. The red colour represents direct occurrence with the source or target entity in the same sentence, the black line represents indirect co-occurrence, and the green line represents potential co-occurrence.

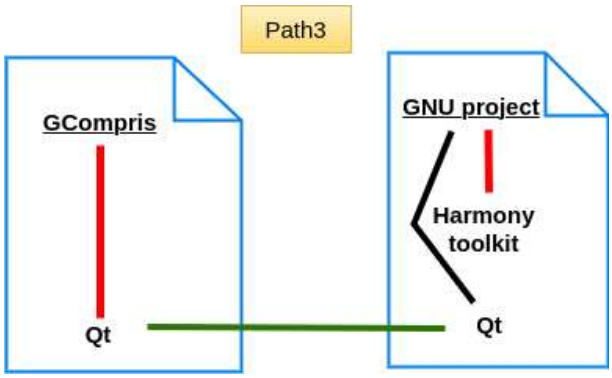

Figure 6.4: An example of a co-occurring graph for path 3 in Figure 6.1.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

For a mentioned entity e^m in each text path p_i, the score of each mentioned entity e^m is calculated by the following Equation,

$$\text{score}(e^m) = \lambda S_1(e^m) + \eta S_2(e^m) + \kappa S_3(e^m)$$

$$S_1(e^m) = \begin{cases} 1, & \text{if } \Theta_1(e^m) \\ 0, & \text{otherwise} \end{cases}$$

$$S_2(e^m) = \begin{cases} \left|\{e^o \mid (\Theta_1(e^o) \wedge I(e^o)) = 1\}\right|, & \text{if } \Theta_2(e^m) \\ 0, & \text{otherwise} \end{cases}$$

$$S_3(e^m) = \begin{cases} \left|\{p_i \mid e^m \in \mathrm{E}_i^m\}\right|, & \text{if } \Theta_3(e^m) \\ 0, & \text{otherwise} \end{cases}$$

where λ, η, κ are hyper parameters and $I(e^o) = 1$ when e^o and e^m are co-occuring in the same sentence, where $e^o \in E_i^m \backslash \{e^m\}$. S_2 sums the number of occurrences of e^o and S_3 sums the occurrence of mentioned entity in path p_i. Next, we calculate the importance score Imp^s of each sentence by aggregating the scores of each mentioned entity, $Imp^s = \sum_{e^m \in E_s^m} \text{score}(e^m)$, where E_s^m is mentioned entity (bridge) in sentence s. The sentences are then ranked based on their importance score, and the top K sentences form the candidate set, $S = \{s_1, s_2, ..., s_K\}$, where K is set to 16 based on the experiment results. We reused the entity based filter from ECRIM model [102].

6.2.4 Relevance-based Filter

In this step, we aim to refine the candidate set of sentences obtained from the previous stages by incorporating additional context derived from entity type information and connecting context. This process ensures that the most semantically relevant and informative sentences are selected for further analysis.

We begin by aggregating the candidate set sentences with the context selected using entity type information. This aggregated set is denoted as S_{total}, which is the union of the original candidate set S and the domain knowledge obtained from the Entity Connecting Context (ECC), represented as S_{ECC}. Mathematically, this can be expressed as:

$$S_{total} = S \cup \{S_{ECC}\}$$

The next step involves applying a relevance-based filter to S_{total}. This filter evaluates the semantic relevance of each sentence within the set. The underlying assumption here is that sentences which are semantically similar to each other, particularly those that include the target entity, are

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

likely to be more informative. This similarity indicates a higher potential for containing relevant information about the relationships between entities.

To achieve this, we employ semantic similarity measures, to quantify the relevance of each sentence. These measures help in identifying sentences that are closely related in meaning and context.

Our ultimate objective is to extract the most informative context I^* from the candidate set S_{total}. This context I^* will be used for reasoning about the relations between entities, thereby enhancing the accuracy and depth of our analysis. By focusing on the most relevant and informative sentences, we can improve the overall quality of the information used for subsequent reasoning tasks.

6.2.5 Encoder

We mark the start and end of every entity and context using special tokens in sentences. Following the baseline [52], we have used the BERT model [187] to encode the tokens selected from the previous step. Here, $start$ and end are the start and end positions of the j-th mention, e_j is the j-th mentioned entity, and n is the total number of sentences.

$$e_j = BERT(\{({S_{total}}_{i=1}^{n}\})_{start}^{end}$$

6.2.6 Relation matrix

Besides the mentioned entity, common relations also exist across various text paths. To capture these details from the text path, we used the cross-path entity relation attention module based on Transformer [236]. We collect all entity mentioned representation in a bag and then generate relation representations for entity pairs (e_u, e_v). Here, $\boldsymbol{E}_r$, $\boldsymbol{E}_u$, $\boldsymbol{E}_v$ are learnable parameters and e_u, e_v are combinations of all entities present, including the mentioned entity.

$$\boldsymbol{r}_{u,v} = \mathrm{ReLU}\left(\boldsymbol{E}_r\left(\mathrm{ReLU}\left(\boldsymbol{E}_u\boldsymbol{e}_u + \boldsymbol{E}_v\boldsymbol{e}_v\right)\right)\right)$$

To model the relation interaction across paths, we build a relation matrix $M \in \mathcal{R}^{|E|\times|E|\times d}$, where $E = \bigcup_{i=1}^{N} E_i$ denotes all the entities in the entity set E_i of text path p_i and $E_i = \{e_s, e_o\} \cup E_i^m$ and N is total number of entities.

6.2.7 Transformer

For capturing the intra and inter path dependencies, we apply a multi-layer Transformer to perform self-attention on the flattened relation matrix $\hat{\boldsymbol{M}} \in \mathbb{R}^{|E|^2\times d}$:

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

$$\hat{\boldsymbol{M}}^{(t+1|1)} = \text{Transformer}(\hat{\boldsymbol{M}}^{(t)})$$

We obtain the target relation representation r_{h_i,t_i} for each path p_i from the last layer of the Transformer.

6.2.8 Classifier

After getting the relation representation r_{h_i,t_i} for each text path p_i for each pair of entities, r_{h_i,t_i} is used as a classification feature. We feed these features to the MLP classifier to get the score for each relation. The relation that gets the maximum score is the predicted relation.

$$\hat{y}_i = \boldsymbol{MLP}(r_{h_i,t_i})$$

A max pooling operation is applied to each relation label to obtain the final score for each relation type r.

$$\hat{y}^{(r)} = \text{Max}\left\{\hat{y}_i^{(r)}\right\}_{i=1}^{N}$$

After obtaining the scores for all relations, a global threshold θ is applied to filter out the categories lower than the threshold. A additional threshold is introduced by baseline paper [102] to control which class should be output. The scores of target classes should be greater than threshold and scores of non target class are less than threshold. Formally, for each Bag B, loss is defined as:

$$\mathcal{L} = \log\left(e^{\theta} + \sum_{r\in\Omega^B_{\text{neg}}} e^{\hat{\boldsymbol{y}}^{(r)}}\right) + \log\left(e^{-\theta} + \sum_{r\in\Omega^B_{\text{pos}}} e^{\hat{\boldsymbol{y}}^{(r)}}\right)$$

Here, $\hat{y_r}$ is score for relation r , θ represent threshold and is set to zero, Ω^B_{pos} and Ω^B_{neg} are positive and negative classes between target entity pair.

6.2.9 Explanation

To improve the interpretability of our model, we incorporated an explainable module that can explain the predicted relationship by providing the filtered sentences given to the model. This represents a notable advancement, as contemporary state-of-the-art models cannot often furnish explanations alongside their predictions. In CrossDocRE, the most challenging issue lies in the length of the documents because it affects the amount of noise the model has to handle. So having an explanation module also helps in getting to know how well the model is able to handle noise. Along with that, it also helps in facilitating error analysis. We retrieve the tokens that are fed to the model (I^*), converting them into strings. This way, we get the exact token data that was used to make the relation prediction. This process enables us to obtain the precise

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

token data, which drives predictions.

6.3 Experimental setup

Metrics. We used F1 and area under curve (AUC) scores following the baseline [52] for a fair comparison. For the open setting, we retrieve the top k document paths from the Wikipedia corpus[1] and use the models trained in the closed setting to predict a relation. These paths are scored using three conditions. 1) Entity count: Entity count refers to the number of times the source entity appears in the document containing the source, and similarly, the number of times the target entity appears in the document containing the target 2) Shared entity: the number of shared entities that appear in source and target documents 3) TF-IDF: The TF-IDF similarity between the two documents. Similar to baseline, we selected the top 16 paths with the highest scores. Hyper-parameters used in our model are shown in Table 6.1.

Hyper parameters	**Value**
Learning rate	3e-5
Embedding dimesion	768
Encoder layers	3
γ, η, κ	0.1,0.01, 0.001
Optimizer	AdamW

Table 6.1: Hyper-parameters setting

Baseline model for comparison. We used all the baseline models available in CrossDocRE for comparison. 1) CodRED [52] is used, which extracts text snippets surrounding the two entities in the document as input and feeds it into a BERT-based model 2) ECRIM [102]; an entity-based document filter is constructed and then fed into a BERT-based model. 3) MR.COD [237] is a multi-hop evidence retrieval method based on evidence path mining and ranking. 4) LGCR [103] discusses local to global reasoning method which enables efficient distinguishing, filtering and global reasoning on complex information from a causal perspective.

We conducted our evaluation in response to the following research questions.

- **RQ1**: What is the effectiveness of KXDocRE in combining context knowledge with reasoning in solving CrossDocRE?

[1]We follow the baseline paper and set k to 16 [52]

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

- **RQ2**: Does the explanation generated by our approach provide sufficient grounds to support the inferred relation?

Dataset. We used the CodRED dataset for evaluation. CodRED contains 11,971 entities and 276 relation types in this dataset. The details of CoDRED dataset are given in Table 6.2. Bags represent relational facts of dataset, 2733 are positive labelled facts and 16,668 are labelled as NA in training set. For more information about CodRED datasets refer to chapter 2.

		Train	Dev	Test
Bags	Pos	2733	1010	1012
	NA	16,668	4558	4523
Text paths	-	129,548	40,740	40,524

Table 6.2: Statistics of CoDRED dataset

6.4 Results

Our model works in the open and closed settings of CrossDocRE. We evaluated our model with three variations: using entity context (EC), connecting context (CC), and both (ECC). We used an NVIDIA A100-SXM4 tensor core GPU with 40GB of memory on Linux 5.4.0-125 with Python version 3.8.5. The results for closed and open settings are available in Table 6.3 and Table 6.4. The results on the test set were obtained from CodaLab[2]. In contrast to a closed setting, performance declines in an open setting due to the retrieval of paths, not all of which significantly contribute to reasoning. Compared to the baseline model, KXDocRE improves by $\approx 3\%$ in the F1 score (closed setting) and $\approx 4\%$ in the F1 score (open setting). In all settings of KXDocRE, it was able to reason over text better than baseline. Our findings demonstrated the significant role domain knowledge can play in the process of reasoning. Hence, our to answer **RQ1** is affirmative. Apart from BERT, we used GPT2[3] in our evaluations. However, there was no improvement in the F1 score.

[2] https://codalab.lisn.upsaclay.fr/competitions/3770
[3] https://huggingface.co/docs/transformers/en/model_doc/gpt2

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

Closed setting					
Base model	**PLM/GNN**	**Dev**		**Test**	
		F1	**AUC**	**F1**	**AUC**
CoDRED	BERT	51.26	47.94	51.02	47.46
ECRIM	BERT	61.12	60.91	62.48	60.67
MR.COD	BERT	61.20	59.22	62.53	61.68
LGCR	BERT	61.67	63.17	61.08	60.75
LGCR	RoBERTa	63.18	64.76	63.79	63.03
KXDocRE$_{ECC}$	**GPT-2**	60.20	59.50	61.80	60.90
KXDocRE$_{EC}$	**BERT**	**63.57**	**62.80**	**65.30**	**64.45**
KXDocRE$_{CC}$	**BERT**	**64.00**	**63.70**	**65.80**	**64.90**
KXDocRE$_{ECC}$	**BERT**	**64.97**	**64.30**	**66.30**	**65.55**

Table 6.3: Results on CodRED dataset for closed setting

Open setting					
Base model	**PLM/GNN**	**Dev**		**Test**	
		F1	**AUC**	**F1**	**AUC**
CoDRED	BERT	47.23	40.86	45.06	39.05
MR.COD	BERT	53.06	51.00	57.88	53.30
LGCR	BERT	52.96	51.48	53.45	50.15
LGCR	RoBERTa	55.15	52.36	55.37	49.05
KXDocRE$_{ECC}$	**GPT-2**	54.20	53.50	55.80	54.90
KXDocRE$_{EC}$	**BERT**	**55.50**	**54.30**	**56.15**	**50.11**
KXDocRE$_{CC}$	**BERT**	**55.90**	**54.80**	**57.12**	**50.60**
KXDocRE$_{ECC}$	**BERT**	**56.70**	**55.20**	**57.93**	**57.12**

Table 6.4: Results on CodRED dataset for open setting

6.4.1 Ablation Study

Effectiveness of relevance and entity-based filter: We studied the impact of relevance and entity-based filters on the performance of KXDocRE. Figure 6.5 shows the F1 score obtained using the Dev dataset. After removing the relevance filter, the performance dropped by 7.7% and if the entity-based filter is removed, the performance dropped by 4.9%. After removing both filters, the F1 score drops by 12%. This indicates that the two filters play an important role in KXDocRE.

Effectiveness of explanations: In cross-document setting, predicting the relationship between entity pairs involves considering multiple text paths (averaging 3646) found across various doc-

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

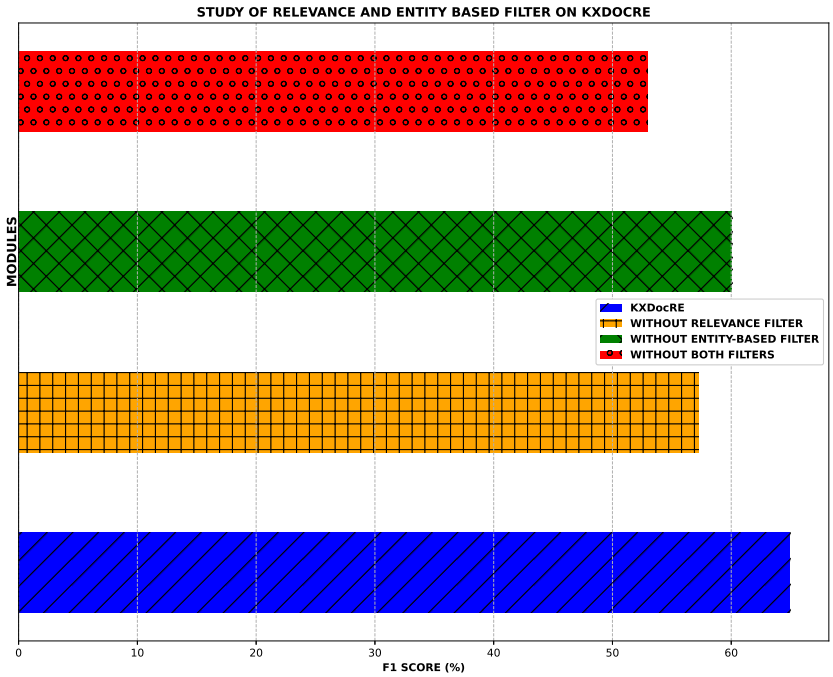

Figure 6.5: Study of relevance and entity based filter on KXDocRE

uments. The explanation module aids in surfacing the most relevant paths that influence the prediction. This module is expected to help us to understand the prediction using a single paragraph containing the relevant paths, rather than scrutinizing each document individually. We discuss the explanation module using a case study (Figure 6.6). The labelled relation between Oichi (Q635214) and Ohatsu (Q1050395) is *Child (P40)*. The explanation text provides text paths between Oichu and Ohatsu via Azai, Toyotomi and Yodo-dono (Oichi is the spouse of Azai Nagamasa, who is the child of Yodo-dono, who is the sibling of Ohatsu); hence, Oichi is the child of Ohatshu. Therefore, the explanation text provides enough evidence along with the context to reason over the data to understand the predicted relationship. Hence, we answer **RQ2**.

Effectiveness of number of hops on KXDocRE: We also studied the impact of the number of hops considered for extracting context in a CC setting on the F1 score (Figure 6.8). The F1 score increases with the number of hops until a point, and after that, the F1 score starts decreasing and saturates. Increasing the hops does not add much new and relevant information after a certain point. Due to this reason, we considered the number of hops (N_h) up to 5.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

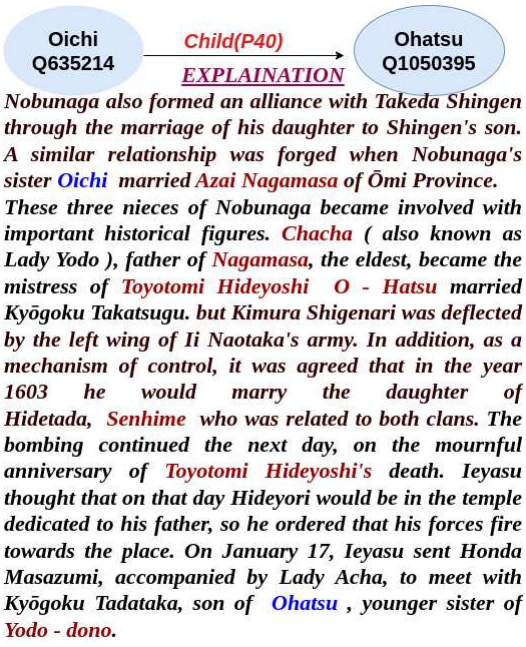

Figure 6.6: Explanation generated using KXDocRE. The source and target entities are blue in colour and intermediate entities are in red color.

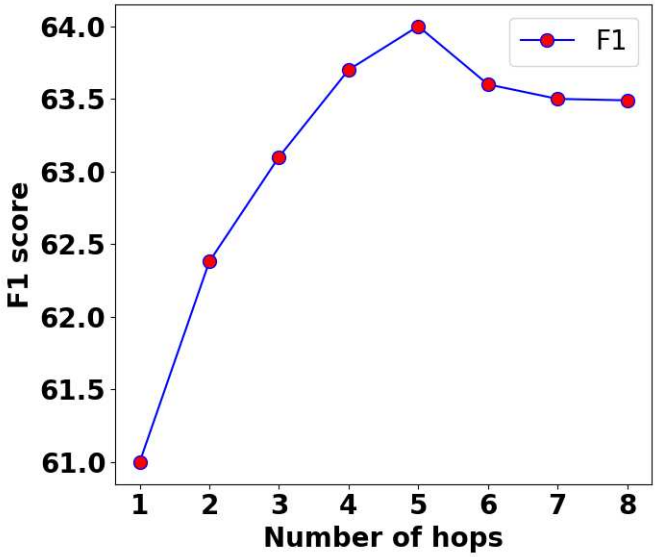

Figure 6.8: Effectiveness of hops in KXDocRE

Error analysis: We studied the successful and failed cases of KXDocRE based on the domain knowledge of the entity pair. From Table 6.5, we can say that the likelihood of a correct prediction is higher if context is available for the given entity pair.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

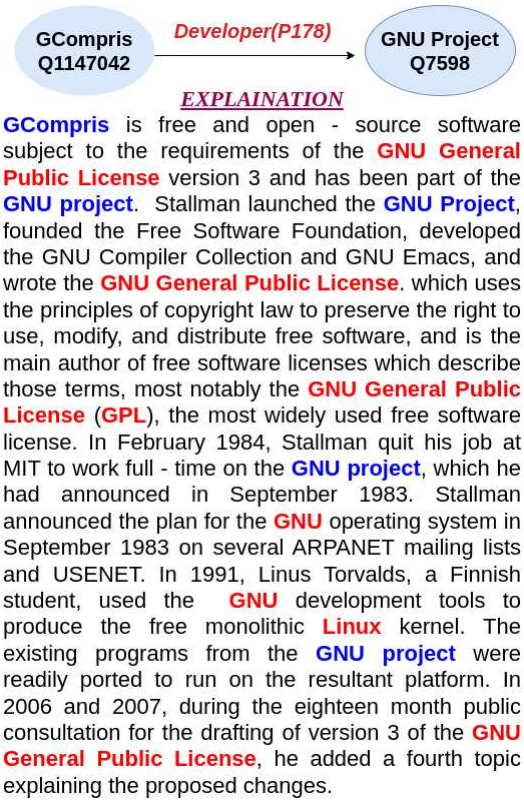

Figure 6.7: Explanation generated from KXDocRE for example discussed in Figure 6.1

	# Dev entity pair	**# Correct prediction**	**# Incorrect prediction**	**Correct prediction%**
KXDocRE$_{EC}$	952	894	58	93.9
KXDocRE$_{CC}$	921	890	31	96.6

Table 6.5: Impact of context in successful cases on Dev dataset

Complexity analysis: In the CodRED dataset, the text path consists of 129,548 instances in the training set, 40,740 instances in the development set, and 40,524 instances in the test set. The average time of a single epoch's execution in the baseline and our model is given in Table 6.6. The execution time for KXDocRE is longer (linear increment) due to the addition of context in the module. The time taken to create domain knowledge is available in Table 6.7.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

Model	Total Time (in hours)	Total (#epochs)	Average time (in seconds) per epoch
ECRIM (baseline)	76	10	7.6
KXDocRE(ours)	96.8	10	9.6

Table 6.6: Complexity analysis of KXDocRE as compared to baseline.

Module	Average time (in sec)
Entity context	0.001
Context path (1-hop)	0.11
Context path (2-hop)	0.21
Context path (3-hop)	0.38

Table 6.7: Average time taken to create the context in KXDocRE.

Case study: We discuss two successful cases and one failed case of KXDocRE and compared them with the baseline model (Table 6.8). **Case1:** To identify the relation between *Dreamlover (Q909801)* and *If it's Over (Q1095958)* from Documents 1 and 2, we use domain knowledge of both entities. EC does not exist for the given entity pair, hence we use CC. CC for both entities is {*part of, Emotions, tracklist, If It's Over, followed by*}. Adding CC helps KXDocRE to predict the correct relation compared to our baseline model, ECRIM. **Case2:** In second case study, for entity pairs *Airbus A320neo family (Q6488)* and *Airbus (Q67)*, we add the ECC context as {*owned by, ORG, ORG*}, which helps KXDocRE help in predicting the relation. **Case3:** We studied a failed case of KXDocRE for entity pairs *Adium (Q58058)* and *x86_64 (Q272629)*. CC for this entity pair is {*instance of, free software, subclass of, software, model item, Mozilla Firefox model item*}. This context does not contribute significantly in predicting the relationship.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

Case1:

Document 1: Dreamlover is a song by American singer Mariah Carey ....Dreamlover marked a more pronounced attempt on Carey's part to incorporate .... arey began to alter her songwriting style and genre choices, most notably in Dreamlover. Dreamlover an Vision of Love Carey's best, calling them the original hits.....

Document 2: If It's Over If It's Over is a song written by American singers and songwriters Mariah Carey and .....if it's over, let me go Several months after the release of Emotions Carey performed the song...If It's Over is a downtempo ballad, which incorporates several genres.....

Correct answer: **followed by**

CodRED: N/A

ECRIM: N/A

KXDocRE$_{ECC}$: followed by

Case2:

Document 1: It is the core development area of of Bandai Namco group. .....main video game branch of Bandai Namco Holdings. In Feb 2005, primariliy set in fictional Japanese city of ....., in association with the Japanese government, suppressed information can be.....

Document 2: Bandai Visual, Bandai Entertainment, Dentsu, Nippon.....and original Japanese (one late-night screening).....It was released on 4 March 2004 in Japan and 8 November 2004 in North America.....The second volume Ghost in the shell.....in Japan and on 26 September 2006.....

Correct answer: **parent organization**

CodRED: N/A

ECRIM: N/A

KXDocRE$_{ECC}$:parent organization

Case3:

Document 1: Adium is a free and open source instant messaging client for macOS that supports... including Windows Live Messenger...It is written using macOS...under the GNU....Adium makes use of a plugin architecture...

Document 2: In computer architecture, 64-bit integers, memory.....ALU architectures are those that are based on.....AMD released its first x86-64...Java program can run on a 32- or ...

Correct answer: **N/A**

CodRED: N/A

ECRIM: N/A

KXDocRE$_{ECC}$:N/A

Table 6.8: Case study

6.5 Summary

In this chapter, we introduce a novel model, KXDocRE, with three settings (EC, CC, ECC) that incorporate domain knowledge in CrossDocRE. These settings represent different types of context, corresponding to entity type, connecting path, and a combination of entity type with connecting path, which are utilized in our framework. Our results suggest that model performance is enhanced by integrating diverse domain knowledge across these variations. Specifically, our model improves the F1 score by approximately 3% in the closed setting and 4% in the open setting. Additionally, we provide explanatory text for each prediction, making our model more interpretable. For future work, researchers can develop a versatile model capable

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

of handling diverse document sets and accumulating domain knowledge. The source code of KXDocRE is available under the Apache License, Version 2.0, at `https://github.com/kracr/cross-doc-relation-extraction`.

6.6 Limitations

Our work has certain limitations when it comes to generating connecting context. This context is only established if a path exists between two entities in Wikidata. To address this limitation, we aim to integrate additional external knowledge bases to enhance context creation. However, as the number of text paths and referenced entities grows, it leads to higher GPU memory usage and slower processing speeds.

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

Algorithm 2: ContextGeneration

Input : Entity pair (v_s,v_o), Number of hops (N_h)
Output: Context

Initialization: i ← 1, source ← v_s, $path_{i-1}$, finalcontext, finalentitytype, adjacent_node, hop_path_i ← **None**
Context ← ContextGeneration(v_s, v_o, N_h)
Function `ContextGeneration` (v_s, v_o, N_h) **:**
    contextpath, entitytype ← {}
    foreach *entity pair* v_s, v_o ∈ *KnowledgeBase* **do**
        contextpath.append(ExplorePath(v_s, v_o, N_h))
    foreach *entity pair* v_s, v_o **do**
        entitytype.append(ExploreEntityType(v_s, v_o)
    finalcontext ← contextpath ∪ entitytype
    return {finalcontext}
Function `ExplorePath`(v_s, v_o, N_h)**:**
    $edge_i$, adjacent $node_i$ = GetOneHopFromSource(v_s)
    $path_i$ = {$edge_i$, adjacent $node_i$}
    i ← 1
    while $v_o \neq$ *adjacent node and* $i \leq N_h$ **do**
        $path_i$ = $path_i$ ∪ GetOneHopFromSource(adjacent node)
        i ← i + 1
    if *adjacent node* == v_o **then**
        return {$path_i$}
    else
        return {}
Function `ExploreEntityType`(v_s, v_o)**:**
    $entity\_type_s$, $entity\_type_o$ ← {}
    if *EntityTypeExist*(v_s) **then**
        $entity\_type_s$ ← GetEntityType(v_s)
    if *EntityTypeExist*(v_o) **then**
        $entity\_type_o$ ← GetEntityType(v_o)
    finalentitytype ← $entity\_type_s$ ∪ $entity\_type_o$ **return** {finalentitytype}

The work discussed in this chapter is part of the following publication. Monika Jain, Raghava Mutharaju, Kuldeep Singh and Ramakanth Kavuluru. 2024: Knowledge-Driven Cross-Document Relation Extraction. Research Track: The 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Findings, Bangkok, Thailand, August 11–16, 2024.

CHAPTER 7

Conclusion and Future work

Relation extraction is a crucial task in natural language processing (NLP) as it involves identifying and classifying semantic relationships between entities mentioned in the text. This process is essential for various applications, including information retrieval, knowledge graph construction, question answering, and text summarization. By accurately extracting relationships, we can transform unstructured text into structured data, enabling more efficient and effective data analysis and decision-making. The existing approaches employ various techniques such as transformer models [18], graph neural network models [19, 20] and reasoning models [93]. These models are effective but do not consider the background knowledge of entities or existing relations between the entities. In this thesis, we explore the importance of incorporating domain knowledge into relation extraction systems. We highlight the benefits of integrating this knowledge to improve system performance. Our motivation arises from the observation that existing relation extraction systems predominantly rely on explicit evidence from input text to identify and describe semantic relationships between specific concepts. While we recognize the importance of extracting explicit assertions to build knowledge bases where they do not exist, we argue that enhancing relational extraction systems with one or multiple sources of background knowledge can significantly enrich both the depth of system inputs and the context of final outputs. By carefully balancing the use of existing knowledge bases with the extraction of new information, our approach aims to mitigate potential limitations in generalizability while still leveraging the valuable insights provided by established knowledge sources. Our approach comprises of three levels of granularity. In the first level of granularity, we leverage expressive axioms from ontology for reasoning in relation extraction and augment F1 score by integrating them with a Graph Neural Network (GNN) model tailored for biomedical datasets. This model, referred to as ReOnto. Evaluation results on two public biomedical datasets, BioRel and ADE, show that ReOnto method performs better than the baselines. ReOnto model is available under the Apache License, Version 2.0, at `https://github.com/kracr/reonto-relation-extraction`.

In the second level of granularity, we subsequently expand our RE problem to encompass document-level relation extraction (DocRE) using a broader dataset of DocRED, ReDocRE and DWIE. We re-frame the DocRE problem as a link prediction task, incorporating entity context from Wikidata and enhancing model interpretability. This model is referred to as DocRE-Clip. Our framework has superior F1 score over baselines on all three datasets. DocREClip

is available under the Apache License, Version 2.0, at https://github.com/kracr/document-level-relation-extraction.

The third level of granularity that we consider is extracting relations from across the documents. We name this approach KXDocRE. It involves incorporating entity context from various paths within Wikidata and feeding them into a model. We introduce three variations of KXDocRE called entity type, connecting path and entity type with connecting path context. In all three variations, KXDocRE performs better than the baseline models. The code of KXDocRE is available under the Apache License, Version 2.0, at https://github.com/kracr/cross-doc-relation-extraction.

While these three models share the common goal of extracting relationships between entities, their architectures and methodologies are tailored to the specific challenges of each task. ReOnto focuses on sentence-level relation extraction, where the relationship between entities is inferred within the boundaries of a single sentence. Its architecture is optimized for handling sentence-level dependencies and symbolic knowledge from ontologies. DocREClip is specifically designed for document-level relation extraction, which involves identifying relationships between entities across multiple sentences within a single document. To address the complexity of capturing long-range dependencies and performing logical reasoning throughout the document, we reformulated the DocRE task as a link prediction problem. This approach allows us to leverage graph-based methods to model entity relationships effectively. Additionally, we incorporated a reasoning module to enhance the model's ability to infer relationships by combining contextual information and logical reasoning.

KXDocRE addresses the challenge of cross-document relation extraction, where entities and their relationships are spread across multiple documents. Since entities can appear in various documents, numerous text paths are generated through these mentioned entities. The key challenge lies in identifying and filtering out irrelevant text paths to focus on the ones that are meaningful and accurate. To tackle this, we incorporated filtering mechanisms to extract only the important text paths, ensuring the model processes relevant information effectively.

While these models are built on different architectures, they all share a common approach of incorporating domain knowledge to enhance performance. For ReOnto, this is achieved through the use of ontologies, while for DocREClip and KXDocRE, domain knowledge is integrated using the Wikidata knowledge base. Across all models, we utilize entity details and contextual paths between entities to enrich the relation extraction process.

While we experimented with applying ReOnto to DocRE tasks, the results were not satisfactory, highlighting the limitations of using a sentence-level model for broader contexts. As part of fu-

ture research, we suggest exploring the use of DocREClip for sentence-level relation extraction to assess its adaptability to more localized contexts. Similarly, DocREClip could be extended to CrossDocRE by incorporating a filtering module to handle the complexity of text paths across multiple documents. These directions offer promising opportunities for researchers to further refine and expand the scope of relation extraction models. While we talked about all advantages of these models our work has limitations in terms of creating a context. Such a context will only be created if there is some entity information or path between two entities in Wikidata or ontology. We plan to overcome this limitation by incorporating other external knowledge bases for creating context. As a future work, researchers can create a versatile model that can work on diverse document sets, accumulating domain knowledge. The required domain knowledge can be gathered from ontology or publicly available knowledge bases. This versatile model can consider the various types of input text, such as sentences, documents and cross-document and gather domain knowledge to predict the relation.

REFERENCES

[1] Guus Schreiber and Yves Raimond. Rdf 1.1 primer. `https://www.w3.org/TR/rdf11-primer/`, 2014. World Wide Web Consortium (W3C) Recommendation.

[2] OpenGenus Foundation. Graph neural networks, 2023. Accessed: 2023-10-01.

[3] Kyle Hamilton, Aparna Nayak, Bojan Bozic, and Luca Longo. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, 09 2022.

[4] Kartik Detroja, C.K. Bhensdadia, and Brijesh S. Bhatt. A survey on relation extraction. *Intelligent Systems with Applications*, 19:200244, 2023.

[5] Wenjing Wu, Caifeng Wen, Qi Yuan, Qiulan Chen, and Yunzhong Cao. Construction and application of knowledge graph for construction accidents based on deep learning. *Engineering, Construction and Architectural Management*, 32(2):1097–1121, 2025.

[6] Somayyeh Behmanesh, Alireza Talebpour, Mehrnoush Shamsfard, and Mohammad Mahdi Jafari. Improved relation span detection in question answering systems over extracted knowledge bases. *Expert Systems with Applications*, 224:119973, 2023.

[7] Lochan Basyal and Mihir Sanghvi. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models, 2023.

[8] Yanying Mao, Qun Liu, and Yu Zhang. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences*, 36(4):102048, 2024.

[9] Michał Gałusza and Andrzej Walczak. Graph-based document-level relationship extraction for risk analysis: A transitive and dialog coherence approach. *Expert Systems with Applications*, 257:124990, 2024.

[10] Richard C. Mohs and Nigel H. Greig. Drug discovery and development: Role of basic biological research. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*, 3(4):651–657, Nov 2017.

[11] Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. Progress in machine translation. *Engineering*, 18:143–153, 2022.

[12] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July 2019. Association for Computational Linguistics.

[13] Haiyun Jiang, Qiaoben Bao, Qiao Cheng, Deqing Yang, Li Wang, and Yanghua Xiao. Complex relation extraction: Challenges and opportunities, 2020.

[14] Kartik Detroja, C.K. Bhensdadia, and Brijesh S. Bhatt. A survey on relation extraction. *Intelligent Systems with Applications*, 19:200244, 2023.

[15] Monika Jain, Kuldeep Singh, and Raghava Mutharaju. Reonto: A neuro-symbolic approach for biomedical relation extraction. In *Machine Learning and Knowledge Discovery in Databases: Research Track: European Conference, ECML PKDD 2023, Turin, Italy, September 18–22, 2023, Proceedings, Part IV*, page 230–247, Berlin, Heidelberg, 2023. Springer-Verlag.

[16] Monika Jain. Knowledge enabled relation extraction. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1210–1213, New York, NY, USA, 2024. Association for Computing Machinery.

[17] Phil Crone. Deeper task-specificity improves joint entity and relation extraction. *arXiv preprint arXiv:2002.06424*, 2020.

[18] Markus Eberts and Adrian Ulges. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020*, pages 2006–2013. IOS Press, 2020.

[19] Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. Recon: relation extraction using knowledge graph context in a graph neural network. In *Proceedings of the Web Conference 2021*, pages 1673–1685, 2021.

[20] Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1331–1339, Florence, Italy, July 2019. Association for Computational Linguistics.

[21] Abhishek Nadgeri, Anson Bastos, Kuldeep Singh, Isaiah Onando Mulang', Johannes Hoffart, Saeedeh Shekarpour, and Vijay Saraswat. KGPool: Dynamic knowledge graph context selection for relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 535–548, Online, August 2021. Association for Computational Linguistics.

[22] Santosh T.Y.S.S, Prantika Chakraborty, Sudakshina Dutta, Debarshi Kumar Sanyal, and Partha Pratim Das. Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types. In *EEKE@JCDL*, 2021.

[23] Muhammet Balcilar, Guillaume Renton, Pierre Héroux, Benoit Gaüzère, Sébastien Adam, and Paul Honeine. Analyzing the expressive power of graph neural networks in a spectral perspective. In *International Conference on Learning Representations*, 2020.

[24] Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Hiroki Kanezashi, Toyotaro Suzumura, and Isaiah Onando Mulang'. How expressive are transformers in spectral domain for graphs? *Transactions on Machine Learning Research*, 2022.

[25] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledge base. *Communications of the ACM*, 57:78–85, 2014.

[26] Jia-Fei Hong and Chu-Ren Huang. Ontology-based event relation prediction: A sumo based study of mandarin vv compounds. *Journal of Chinese Linguistics*, 43:170–195, 01 2015.

[27] Patricia L. Whetzel, Natasha Noy, Nigam Haresh Shah, Paul R. Alexander, Csongor Nyulas, Tania Tudorache, and Mark A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39:W541 – W545, 2011.

[28] Jonathan P. Bona, Mathias Brochhausen, and William R. Hogan. Enhancing the drug ontology with semantically-rich representations of national drug codes and RxNorm unique concept identifiers. *BMC Bioinformatics*, 20(21), December 2019.

[29] Rainer Winnenburg, Jonathan M Mortensen, and Olivier Bodenreider. Using description logics to evaluate the consistency of drug-class membership relations in NDF-RT. *Journal of Biomedical Semantics*, 6:13, 2015.

[30] Jeff Z Pan, Mei Zhang, Kuldeep Singh, Frank van Harmelen, Jinguang Gu, and Zhi Zhang. Entity enabled relation linking. In *International Semantic Web Conference*, pages 523–538. Springer, 2019.

[31] Wen Zhang, Jiaoyan Chen, Juan Li, Zezhong Xu, Jeff Z Pan, and Huajun Chen. Knowledge graph reasoning with logics and embeddings: Survey and perspective. *arXiv preprint arXiv:2202.07412*, 2022.

[32] Jia-Fei Hong, Xiang-Bing Li, and Chu-Ren Huang. Ontology-based prediction of compound relations : A study based on SUMO. In *Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation*, pages 151–160. Logico-Linguistic Society of Japan, dec 2004.

[33] Blend Berisha and Endrit Mëziu. Big data analytics in cloud computing: An overview. *Journal of Cloud Computing*, 02 2021.

[34] Sarah Alanazi, Nazar Mohamed, Mutsam Jarajreh, and Saad Algarni. Question answering systems: A systematic literature review. *International Journal of Advanced Computer Science and Applications*, 12, 01 2021.

[35] Tong Shen, Fu Zhang, and Jingwei Cheng. A comprehensive overview of knowledge graph completion. *Knowledge-Based Systems*, 255:109597, 2022.

[36] Steven Burrows, Alexandra L. Uitdenbogerd, and Andrew Turpin. Application of information retrieval techniques for source code authorship attribution. In Xiaofang Zhou, Haruo Yokota, Ke Deng, and Qing Liu, editors, *Database Systems for Advanced Applications*, pages 699–713, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[37] Peng Sun, Xuezhen Yang, Xiaobing Zhao, and Zhijuan Wang. An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278, 2018.

[38] Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. Graph neural networks with generated parameters for relation extraction. In *Proceedings of ACL*, 2019.

[39] Shanchan Wu and Yifan He. Enriching pre-trained language model with entity information for relation classification. In *CIKM '19: Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2361–2364, 11 2019.

[40] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[41] I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wen-Huang Cheng, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. A simple and unified tagging model with priming for relational structure predictions. *ArXiv*, abs/2205.12585, 2022.

[42] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[43] Cícero dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 626–634, Beijing, China, July 2015. Association for Computational Linguistics.

[44] Rui Cai, Xiaodong Zhang, and Houfeng Wang. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765, Berlin, Germany, August 2016. Association for Computational Linguistics.

[45] Yiming Liu, Hongtao Shan, Feng Nie, Gaoyu Zhang, and George Xianzhi Yuan. Document-level relation extraction with local relation and global inference. *Information*, 14(7), 2023.

[46] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In *AAAI Conference on Artificial Intelligence*, 2020.

[47] Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4925–4936, Hong Kong, China, November 2019. Association for Computational Linguistics.

[48] Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online, July 2020. Association for Computational Linguistics.

[49] Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 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, November 2020. Association for Computational Linguistics.

[50] Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, Online, August 2021. Association for Computational Linguistics.

[51] Monika Jain, Raghava Mutharaju, Ramakanth Kavuluru, and Kuldeep Singh. Revisiting document-level relation extraction with context-guided link prediction. In *Proceedings of The 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, 2024. AAAI.

[52] Yuan Yao, Jiaju Du, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. CodRED: A cross-document relation extraction dataset for acquiring knowledge in the wild. In *Proceedings of EMNLP 2021*, pages 4452–4472, 2021.

[53] Nguyen Bach and Sameer Badaskar. A review of relation extraction. *Unpublished manuscript*, 05 2011. Accessed via ResearchGate.

[54] Sachin Pawar, Girish K. Palshikar, and Pushpak Bhattacharyya. Relation extraction : A survey, 2017.

[55] Sandra Collovini, Tiago Bonamigo, and Renata Vieira. A review on relation extraction with an eye on portuguese. *Journal of the Brazilian Computer Society*, 19, 11 2013.

[56] Shantanu Kumar. A survey of deep learning methods for relation extraction, 2017.

[57] Hailin Wang, Ke Qin, Rufai Yusuf, Guoming Lu, and Jin Yin. Deep neural network-based relation extraction: an overview. *Neural Computing and Applications*, 34:1–21, 03 2022.

[58] Zara Nasar, Syed Waqar Jaffry, and Muhammad Malik. Named entity recognition and relation extraction: State of the art. *ACM Computing Surveys*, 54, 02 2021.

[59] Yong Shi, Yang Xiao, and Lingfeng Niu. A brief survey of relation extraction based on distant supervision. In *Computational Science – ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part III*, page 293–303, Berlin, Heidelberg, 2019. Springer-Verlag.

[60] Deyu Zhou and Dayou Zhong. Biomedical relation extraction: From binary to complex. *Computational and mathematical methods in medicine*, 2014:298473, 08 2014.

[61] Nada Boudjellal, Huaping Zhang, Asif Khan, and Arshad Ahmad. Biomedical relation extraction using distant supervision. *Scientific Programming*, 2020:1–9, 06 2020.

[62] Julien Delaunay, Tran Thi Hong Hanh, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. A comprehensive survey of document-level relation extraction (2016-2023). *CoRR*, abs/2309.16396, 2023.

[63] Jiaqi Wang, Yahui Li, Jing Huang, and Wenbin Zhao. A survey of document-level relation extraction. In Yongquan Yan, editor, *International Conference on Computer, Artificial Intelligence, and Control Engineering (CAICE 2022)*, volume 12288, page 122881S. International Society for Optics and Photonics, SPIE, 2022.

[64] Wenxuan Zhou and Muhao Chen. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only, November 2022. Association for Computational Linguistics.

[65] Kamel Nebhi. A rule-based relation extraction system using dbpedia and syntactic parsing. In *Proceedings of the 2013th International Conference on NLP & DBpedia - Volume 1064*, NLP-DBPEDIA'13, page 74–79, Aachen, DEU, 2013. CEUR-WS.org.

[66] Bryan Rink and Sanda Harabagiu. UTD: Classifying semantic relations by combining lexical and semantic resources. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 256–259, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[67] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, page 22–es, USA, 2004. Association for Computational Linguistics.

[68] GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 427–434, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[69] George Miller, R. Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. Introduction to wordnet: An on-line lexical database*. 3, 01 1991.

[70] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3(null):1083–1106, mar 2003.

[71] Razvan Bunescu and Raymond Mooney. A shortest path dependency kernel for relation extraction. In Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff, editors, *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[72] Razvan C. Bunescu and Raymond J. Mooney. Subsequence kernels for relation extraction. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, page 171–178, Cambridge, MA, USA, 2005. MIT Press.

[73] Zhu Zhang. Weakly-supervised relation classification for information extraction. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, page 581–588, New York, NY, USA, 2004. Association for Computing Machinery.

[74] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *J. Mach. Learn. Res.*, 2:419–444, mar 2002.

[75] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 415–422, Barcelona, Spain, July 2004.

[76] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[77] William Hogan. An overview of distant supervision for relation extraction with a focus on denoising and pre-training methods, 2022.

[78] Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in*

Natural Language Processing, pages 1257–1266, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[79] Dèlia Fernàndez-Cañellas, Joan Marco Rimmek, Joan Espadaler, Blai Garolera, Adrià Barja, Marc Codina, Marc Sastre, Xavier Giro-i Nieto, Juan Carlos Riveiro, and Elisenda Bou-Balust. Enhancing online knowledge graph population with semantic knowledge. In Jeff Z. Pan, Valentina Tamma, Claudia d'Amato, Krzysztof Janowicz, Bo Fu, Axel Polleres, Oshani Seneviratne, and Lalana Kagal, editors, *The Semantic Web – ISWC 2020*, pages 183–200, Cham, 2020. Springer International Publishing.

[80] Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. RECON: relation extraction using knowledge graph context in a graph neural network. In Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1673–1685. ACM / IW3C2, 2021.

[81] Yuhao Zhang, Peng Qi, and Christopher D. Manning. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[82] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence n-ary relation extraction with graph LSTMs. *Transactions of the Association for Computational Linguistics*, 5:101–115, 2017.

[83] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 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, July 2019. Association for Computational Linguistics.

[84] Chris Quirk and Hoifung Poon. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain, April 2017. Association for Computational Linguistics.

[85] Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. Learning logic rules for document-level relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

1239–1250, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[86] Quzhe Huang, Shengqi Zhu, Yansong Feng, Yuan Ye, Yuxuan Lai, and Dongyan Zhao. Three sentences are all you need: Local path enhanced document relation extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 998–1004, Online, August 2021. Association for Computational Linguistics.

[87] Zhijiang Guo, Yan Zhang, and Wei Lu. Attention guided graph convolutional networks for relation extraction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 241–251, Florence, Italy, July 2019. Association for Computational Linguistics.

[88] Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. Inter-sentence relation extraction with document-level graph convolutional neural network. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4309–4316, Florence, Italy, July 2019. Association for Computational Linguistics.

[89] Huiwei Zhou, Shixian Ning, Yunlong Yang, Zhuang Liu, Chengkun Lang, and Yingyu Lin. Chemical-induced disease relation extraction with dependency information and prior knowledge. *Journal of Biomedical Informatics*, 84:171–178, 2018.

[90] Angrosh Mandya, Danushka Bollegala, Frans Coenen, and Katie Atkinson. A dataset for inter-sentence relation extraction using distant supervision. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[91] Ye Wu, Ruibang Luo, Henry C. M. Leung, Hing-Fung Ting, and Tak-Wah Lam. Renet: A deep learning approach for extracting gene-disease associations from literature. In Lenore J. Cowen, editor, *Research in Computational Molecular Biology*, pages 272–284, Cham, 2019. Springer International Publishing.

[92] Komal K. Teru, Etienne G. Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9448–9457. PMLR, 2020.

[93] Wang Xu, Kehai Chen, and Tiejun Zhao. Discriminative reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1653–1663, Online, August 2021. Association for Computational Linguistics.

[94] Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. Document-level relation extraction as semantic segmentation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[95] Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14612–14620. AAAI Press, 2021.

[96] Youmi Ma, An Wang, and Naoaki Okazaki. DREEAM: guiding attention with evidence for improving document-level relation extraction. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1963–1975. Association for Computational Linguistics, 2023.

[97] Chaojun Xiao, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Maosong Sun, Fen Lin, and Leyu Lin. Denoising relation extraction from document-level distant supervision. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3683–3688, Online, November 2020. Association for Computational Linguistics.

[98] Junpeng Li, Zixia Jia, and Zilong Zheng. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, Singapore, December 2023. Association for Computational Linguistics.

[99] Severine Verlinden, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. Injecting knowledge base information into end-to-end joint entity and relation extraction and coreference resolution. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1952–1957. Association for Computational Linguistics, 2021.

[100] Xinyi Wang, Zitao Wang, Weijian Sun, and Wei Hu. Enhancing document-level relation extraction by entity knowledge injection. In Ulrike Sattler, Aidan Hogan, C. Maria Keet, Valentina Presutti, João Paulo A. Almeida, Hideaki Takeda, Pierre Monnin, Giuseppe Pirrò, and Claudia d'Amato, editors, *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, volume 13489 of *Lecture Notes in Computer Science*, pages 39–56. Springer, 2022.

[101] Limin Yao, Sebastian Riedel, and Andrew McCallum. Collective cross-document relation extraction without labelled data. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, page 1013–1023, USA, 2010. Association for Computational Linguistics.

[102] Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[103] Haoran Wu, Xiuyi Chen, Zefa Hu, Jing Shi, Shuang Xu, and Bo Xu. Local-to-global causal reasoning for cross-document relation extraction. *IEEE/CAA Journal of Automatica Sinica*, 10(7):1608–1621, 2023.

[104] Limin Yao, Sebastian Riedel, and Andrew McCallum. Collective cross-document relation extraction without labelled data. In Hang Li and Lluís Màrquez, editors, *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1013–1023, Cambridge, MA, October 2010. Association for Computational Linguistics.

[105] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*, pages 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics.

[106] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[107] C. Walker and Linguistic Data Consortium. *ACE 2005 Multilingual Training Corpus*. LDC corpora. Linguistic Data Consortium, 2005.

[108] Evan Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium*, 2008.

[109] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

[110] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. of Biomedical Informatics*, 45(5):885–892, oct 2012.

[111] Olivier Taboureau, Sonny Kim Nielsen, Karine Audouze, Nils Weinhold, Daniel Edsgärd, Francisco S. Roque, Irene Kouskoumvekaki, Alina Bora, Ramona Curpan, Thomas Skøt Jensen, Søren Brunak, and Tudor I. Oprea. Chemprot: a disease chemical biology database. *Nucleic Acids Research*, 39:D367 – D372, 2010.

[112] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016, 2016.

[113] S. Marchesin and G. Silvello. Tbga: A large-scale gene-disease association dataset for biomedical relation extraction, 2022.

[114] Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4):102563, 2021.

[115] Qiao Cheng, Juntao Liu, Xiaoye Qu, Jin Zhao, Jiaqing Liang, Zhefeng Wang, Baoxing Huai, Nicholas Jing Yuan, and Yanghua Xiao. HacRED: A large-scale relation extraction dataset toward hard cases in practical applications. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2819–2831, Online, August 2021. Association for Computational Linguistics.

[116] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting docred – addressing the false negative problem in relation extraction. In *Proceedings of EMNLP*, 2022.

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

[118] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. volume 3408, pages 345–359, 04 2005.

[119] Ellen Riloff. Automatically constructing a dictionary for information extraction tasks. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, AAAI'93, page 811–816. AAAI Press, 1993.

[120] Jerry R. Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyalna, and Mabry Tyson. FASTUS: A system for extracting information from text. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

[121] Eugene Agichtein and Luis Gravano. Snowball: extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, DL '00, page 85–94, New York, NY, USA, 2000. Association for Computing Machinery.

[122] Katrin Fundel, Robert Küffner, and Ralf Zimmer. RelEx—Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 12 2006.

[123] Pere-Lluís Huguet Cabot and Roberto Navigli. 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, November 2021. Association for Computational Linguistics.

[124] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 4904–4917, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[125] Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. UniRel: Unified representation and interaction for joint relational triple extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7087–7099, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[126] Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. Joint entity and relation extraction with set prediction networks. *arXiv preprint arXiv:2011.01675*, 2020.

[127] Amir DN Cohen, Shachar Rosenman, and Yoav Goldberg. Relation classification as two-way span-prediction, 2021.

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

[129] Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. Sais: Supervising and augmenting intermediate steps for document-level relation extraction. In *NAACL*, 2022.

[130] Sefika Efeoglu and Adrian Paschke. Retrieval-augmented generation-based relation extraction, 2024.

[131] John Giorgi, Gary Bader, and Bo Wang. A sequence-to-sequence approach for document-level relation extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 10–25, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[132] Nacime Bouziani, Shubhi Tyagi, Joseph Fisher, Jens Lehmann, and Andrea Pierleoni. Rexel: An end-to-end model for document-level relation extraction and entity linking. In *NAACL 2024*, 2024.

[133] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation, 2022.

[134] Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. Knowledge-driven cross-document relation extraction, 2024.

[135] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Ray-

mond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[136] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[137] Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019.

[138] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.

[139] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[140] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher

Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[141] Ben Wang. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. `https://github.com/kingoflolz/mesh-transformer-jax`, May 2021.

[142] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022.

[143] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models, 2023.

[144] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 02 2023.

[145] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[146] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[147] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code llama: Open foundation models for code, 2024.

[148] Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. Thinking about GPT-3 in-context learning for biomedical IE? think again. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[149] Yu Su Kai Zhang, Bernal Jiménez Gutiérrez. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of ACL*, 2023.

[150] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. Gpt-re: In-context learning for relation extraction using large language models, 2023.

[151] Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12136–12147, Singapore, December 2023. Association for Computational Linguistics.

[152] Bo Li, Dingyao Yu, Wei Ye, Jinglei Zhang, and Shikun Zhang. Sequence generation with label augmentation for relation extraction. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.

[153] Xilai Ma, Jing Li, and Min Zhang. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, Singapore, December 2023. Association for Computational Linguistics.

[154] Sizhe Zhou, Yu Meng, Bowen Jin, and Jiawei Han. Grasping the essentials: Tailoring large language models for zero-shot relation extraction, 2024.

[155] Guozheng Li, Peng Wang, Jiajun Liu, Yikai Guo, Ke Ji, Ziyu Shang, and Zijie Xu. Meta in-context learning makes large language models better zero and few-shot relation extractors, 2024.

[156] Guozheng Li, Peng Wang, Wenjun Ke, Yikai Guo, Ke Ji, Ziyu Shang, Jiajun Liu, and Zijie Xu. Recall, retrieve and reason: Towards better in-context relation extraction, 2024.

[157] Nicola Guarino, Daniel Oberle, and Steffen Staab. What is an ontology? In *Handbook on Ontologies*, 2009.

[158] S. al and Martin Kuiper. The gene ontology resource: 20 years and still going strong. 01 2019.

[159] Ravneet Kaur, Monika Jain, Ryan M. McAdams, Yao Sun, Shubham Gupta, Raghava Mutharaju, Su Jin Cho, Satish Saluja, Jonathan P. Palma, Avneet Kaur, and Harpreet Singh. An ontology and rule-based clinical decision support system for personalized nutrition recommendations in the neonatal intensive care unit. *IEEE Access*, 11:142433–142446, 2023.

[160] C. Brewster and K. O'Hara. Knowledge representation with ontologies: the present and future. *IEEE Intelligent Systems*, 19(1):72–81, 2004.

[161] David Riaño, Francis Real, Joan Albert López-Vallverdú, Fabio Campana, Sara Ercolani, Patrizia Mecocci, Roberta Annicchiarico, and Carlo Caltagirone. An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. *Journal of Biomedical Informatics*, 45(3):429–446, 2012.

[162] Tatiana Erekhinskaya, Matthew Morris, Dmitriy Strebkov, and Dan Moldovan. *Leveraging Ontologies for Natural Language Processing in Enterprise Applications*, pages 79–85. 02 2020.

[163] Nadine Schuurman and Agnieszka Leszczynski. Ontologies for bioinformatics. *Bioinformatics and biology insights*, 2:187–200, 03 2008.

[164] Ying Ding, Dieter Fensel, Michel Klein, Borys Omelayenko, and Ellen Schulten. The role of ontologies in ecommerce. 01 2004.

[165] Kunal Sengupta and Pascal Hitzler. *Web Ontology Language (OWL)*, pages 1–6. Springer New York, New York, NY, 2017.

[166] Christian Bizer, Maria-Esther Vidal, and Michael Weiss. *Resource Description Framework*, pages 3221–3224. Springer New York, New York, NY, 2018.

[167] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. Semantics and complexity of sparql. *ACM Trans. Database Syst.*, 34(3), September 2009.

[168] Mark A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, 2015.

[169] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), July 2021.

[170] Xuejiao Zhao, Huanhuan Chen, Zhenchang Xing, and Chunyan Miao. Brain-inspired search engine assistant based on knowledge graph. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4386–4400, 2023.

[171] Zeinab Shokrzadeh, Mohammad-Reza Feizi-Derakhshi, Mohammad-Ali Balafar, and Jamshid Bagherzadeh Mohasefi. Knowledge graph-based recommendation system enhanced by neural collaborative filtering and knowledge graph embedding. *Ain Shams Engineering Journal*, 15(1):102263, 2024.

[172] Enayat Rajabi, Allu George, and Karishma Kumar. The role of knowledge graphs in chatbots. *The Electronic Library*, 42, 06 2024.

[173] Rose Catherine and William Cohen. Personalized recommendations using knowledge graphs: A probabilistic logic programming approach. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 325–332, New York, NY, USA, 2016. Association for Computing Machinery.

[174] Yachen Tang, Haiyun Han, Xianmao Yu, Jing Zhao, Guangyi Liu, and Longfei Wei. An intelligent question answering system based on power knowledge graph, 2021.

[175] Jeff Z. Pan. Resource description framework. In *Handbook on Ontologies*, 2009.

[176] Abdalsamad Keramatfar, Mohadeseh Rafiee, and Hossein Amirkhani. Graph neural networks: A bibliometrics overview. *Machine Learning with Applications*, 10:100401, 2022.

[177] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr., Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks, 2019.

[178] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018.

[179] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.

[180] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks?, 2019.

[181] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, pages 1024–1034, 2017.

[182] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272, 2017.

[183] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 974–983, 2018.

[184] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

[185] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer, 2018.

[186] Felipe Almeida and Geraldo Xexéo. Word embeddings: A survey, 2023.

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

[188] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, 36, 09 2019.

[189] Zachary Susskind, Bryce Arden, Lizy K. John, Patrick Stockton, and Eugene B. John. Neuro-symbolic ai: An emerging class of ai workloads and their characterization, 2021.

[190] Guillaume Lample and François Charton. Deep learning for symbolic mathematics, 2019.

[191] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, 2019.

[192] Dieter Fensel, Umutcan Şimşek, Kevin Angele, Elwin Huaman, Elias Kärle, Oleksandra Panasiuk, Ioan Toma, Jürgen Umbrich, and Alexander Wahler. Why we need knowledge graphs: Applications. In *Knowledge Graphs*, pages 95–112. Springer, 2020.

[193] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[194] Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. Clustering-based inference for biomedical entity linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608, 2021.

[195] Xin Lv, Yixin Cao, Lei Hou, Juanzi Li, Zhiyuan Liu, Yichi Zhang, and Zelin Dai. Is multi-hop reasoning really explainable? towards benchmarking reasoning interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8899–8911, 2021.

[196] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[197] Rui Xing, Jie Luo, and Tengwei Song. BioRel: towards large-scale biomedical relation extraction. *BMC Bioinformatics*, 21(16):543, 2020.

[198] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885 – 892, 2012. Text Mining and Natural Language Processing in Pharmacogenomics.

[199] Yongqun He, Sirarat Sarntivijai, Yu Lin, Zuoshuang Xiang, Abra Guo, Shelley Zhang, Desikan Jagannathan, Luca Toldo, Cui Tao, and Barry Smith. OAE: The ontology of adverse events. *Journal of Biomedical Semantics*, 5:29, dec 2014.

[200] Jung-Jin Yang. An ontology-based intelligent agent system for semantic search in medicine. In Jaeho Lee and Mike Barley, editors, *Intelligent Agents and Multi-Agent Systems*, pages 182–193. Springer Berlin Heidelberg, 2003.

[201] Anand Kumar and Barry Smith. Oncology ontology in the nci thesaurus. In Silvia Miksch, Jim Hunter, and Elpida T. Keravnou, editors, *Artificial Intelligence in Medicine*, pages 213–220, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[202] Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In *VS@HLT-NAACL*, 2015.

[203] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[204] Daniil Sorokin and Iryna Gurevych. Context-aware representations for knowledge base relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1784–1789, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[205] Trung-Tin Huynh, Yulan He, Alistair Willis, and Stefan M. Rüger. Adverse drug reaction classification with deep neural networks. In *International Conference on Computational Linguistics*, 2016.

[206] Ashish Rawat, Mudasir Ahmad Wani, Mohammed ElAffendi, Ali Shariq Imran, Zenun Kastrati, and Sher Muhammad Daudpota. Drug adverse event detection using text-based convolutional neural networks (textcnn) technique. *Electronics*, 11(20), 2022.

[207] Hasham Ul Haq, Veysel Kocaman, and David Talby. Mining adverse drug reactions from unstructured mediums at scale, 2022. version: 2.

[208] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*, 2017.

[209] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[210] María Herrero-Zazo, Isabel Segura-Bedmar, Janna Hastings, and Paloma Martínez. DINTO: Using OWL ontologies and SWRL rules to infer drug–drug interactions and their mechanisms. *Journal of Chemical Information and Modeling*, 55(8):1698–1707, aug 2015. Publisher: American Chemical Society.

[211] Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–581, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[212] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, sep 2014.

[213] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov 1995.

[214] Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3065–3072. AAAI Press, 2020.

[215] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with path reasoning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(4), mar 2023.

[216] Anastasia Shimorina, Johannes Heinecke, and Frédéric Herledan. Knowledge extraction from texts based on Wikidata. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 297–304, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics.

[217] Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting docred - addressing the false negative problem in relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8472–8487. Association for Computational Linguistics, 2022.

[218] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, page 556–559, New York, NY, USA, 2003. Association for Computing Machinery.

[219] Anson Bastos, Kuldeep Singh, Abhishek Nadgeri, Saeedeh Shekarpour, Isaiah Onando Mulang, and Johannes Hoffart. Hopfe: Knowledge graph representation learning using inverse hopf fibrations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 89–99, New York, NY, USA, 2021. Association for Computing Machinery.

[220] Anson Bastos, Kuldeep Singh, Abhishek Nadgeri, Johannes Hoffart, Manish Singh, and Toyotaro Suzumura. Can persistent homology provide an efficient alternative for evaluation of knowledge graph completion methods? In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2455–2466, New York, NY, USA, 2023. Association for Computing Machinery.

[221] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In Aldo Gangemi, Roberto Navigli, Maria-Esther Vidal, Pascal Hitzler, Raphaël Troncy, Laura Hollink, Anna Tordai, and Mehwish Alam, editors, *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, volume 10843 of *Lecture Notes in Computer Science*, pages 593–607. Springer, 2018.

[222] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[223] Markus Eisenbach, Jannik Lübberstedt, Dustin Aganian, and Horst-Michael Gross. A little bit attention is all you need for person re-identification. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 7598–7605. IEEE, 2023.

[224] Wang Xu, Kehai Chen, and Tiejun Zhao. Document-level relation extraction with reconstruction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14167–14175. AAAI Press, 2021.

[225] Andrea Rossi, Donatella Firmani, Paolo Merialdo, and Tommaso Teofili. Explaining link prediction systems based on knowledge graph embeddings. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, page 2062–2075, New York, NY, USA, 2022. Association for Computing Machinery.

[226] Xi Victoria Lin, Richard Socher, and Caiming Xiong. Multi-hop knowledge graph reasoning with reward shaping. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3243–3253. Association for Computational Linguistics, 2018.

[227] Shuang Zeng, Yuting Wu, and Baobao Chang. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 524–534, Online, August 2021. Association for Computational Linguistics.

[228] Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[229] Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 257–268, Dublin, Ireland, May 2022. Association for Computational Linguistics.

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

[231] Ridong Han, Tao Peng, Benyou Wang, Lu Liu, and Xiang Wan. Document-level relation extraction with relation correlations. *ArXiv*, abs/2212.10171, 2022.

[232] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[233] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2071–2080. JMLR.org, 2016.

[234] Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. Multi-hop evidence retrieval for cross-document relation extraction. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10336–10351. Association for Computational Linguistics, 2023.

[235] Abhishek Nadgeri, Anson Bastos, Kuldeep Singh, Isaiah Onando Mulang, Johannes Hoffart, Saeedeh Shekarpour, and Vijay Saraswat. Kgpool: Dynamic knowledge graph context selection for relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 535–548, 2021.

[236] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[237] K. Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. Multi-hop evidence retrieval for cross-document relation extraction. *ArXiv*, abs/2212.10786, 2022.