Honours Project Dissertation

Event Driven Causal Information Extraction from Text

Nathan Scott – 18913101

19 May 2025

https://github.com/nathanpaulscott/CausalRE/tree/main/CRE%20Model

1 Introduction

Extracting causal pairs from text is a challenging NLP task, primarily due to the myriad ways causal information and relationships can be encoded in natural language. Consider two examples:

1. > The earthquake caused destruction in the city.

   This sentence contains a causal pair: **"earthquake"** => **"destruction in the city"**. Detecting and extracting this causal relationship is achievable with discriminative (non-generative) Information Extraction (IE) models. This is largely because the spans of text are short, clearly defined, and the relationship is explicitly indicated by the verb *"caused"* located between them.

2. > 'Belliqueux' rapidly outran Landolphe's flagship 'Concorde', leaving Landolphe with no option but to surrender without any serious resistance.

   Here, the causal pair is: **"Belliqueux rapidly outran Landolphe's flagship Concorde"** => **"surrender without any serious resistance"**. Detecting and extracting this pair poses greater challenges for discriminative models, due to the longer spans, the increased complexity of the concepts involved, and the more intricate sentence structure. As spans grow longer, span boundaries become harder to define precisely, and conceptual encodings become more abstract. Although an explicit causal signal is provided between the spans, it is not a common simple form as in the first example.

As illustrated above, extracting causal relationships as in Example 2 is more challenging. This dissertation focuses specifically on detecting and extracting such causal pairs using discriminative IE models.

1.1 Motivation

The central motivation of causal relation extraction, as with other branches of IE, lies in converting unstructured textual data into structured representations, such as knowledge graphs. Incorporating causal relations into knowledge graphs enables more precise causal reasoning and analysis. Specifically, this project was motivated by the need to extract causal information from geological survey reports, which are characterized by formal language and complex, domain-specific terminology.

1.1.1 Why employ discriminative models for this task?

Causal relation extraction (CRE) has been studied extensively for decades [1, 2, 3, 4, 5, 6], with models ranging from pattern-based to statistical approaches, including feature-engineered

machine learning and deep learning paradigms. If one aspect can be drawn from prior research, it could be that CRE differs substantially in complexity from standard IE tasks such as Named Entity Recognition and Relation Extraction (NER-RE) or Event Extraction (EE), which typically involve shorter spans and simpler relation structures. Thus it is of interest to review the viability of more recent advancements in deep learning approaches for CRE.

Of particular interest for this work are the two recent transformer-based variants of IE model architectures that have emerged in the last decade, namely discriminative and generative models. These paradigms differ in how they approach information extraction. Discriminative transformer-based models use encoder-based transformers to produce contextualized token representations from input text and then classify tokens or groupings of tokens, such as spans and span-pairs, against supervised labels. In contrast, generative transformer-based models employ decoder-only or encoder-decoder transformers to generate output as a sequence of tokens based on an input consisting of a prompt and data, treating extraction as a conditional text generation task. While both rely on attention mechanisms and benefit from large-scale pretraining, discriminative models are typically smaller, more constrained, and offer greater predictability. Generative models, by comparison, are more flexible but demand significantly more resources and can produce less predictable outputs.

Since the advent of the transformer [7], discriminative transformer-based models have offered superior performance on IE tasks. However, generative models have recently made rapid advances, offering increased flexibility [8] and adaptability as demonstrated by models such as UIE [9] and GoLLIE [10]. It is, however, not yet clear whether discriminative models retain their historical advantage. Despite this, discriminative models currently remain appealing due to their smaller size, computational efficiency, and suitability for deployment in on-premise environments.

1.2 Causal RE and the Long Span Problem

Causal pairs typically differ from standard NER-RE pairs because they are not always short entities. Often, one or both elements of a causal pair encompass complex ideas, described by longer spans of text. Detecting and clearly delineating each span within a causal pair is essential for correctly identifying their causal relationship and thus poses a significant challenge for causal IE tasks.

Longer text spans describe more complex ideas and thus inherently introduce several complicating factors:

- **Ambiguous boundaries:** Span boundaries become increasingly uncertain as textual structures become less clearly defined.
- **Contextual dilution:** Important contextual information tends to become sparse within longer spans.

- **Increased noise:** Irrelevant linguistic elements and noise become more prevalent within longer spans.
- **Co-reference ambiguity:** Co-referential phrases and associated ambiguity become more common and problematic in longer spans.
- **Annotation complexity:** Longer spans significantly complicate annotation tasks for both humans and automated annotation systems.

1.2.1 How Do Humans Identify Causal Pairs?

Human readers typically do not process text in terms of clearly defined long spans. Instead, they read, identify, and mentally combine smaller informational chunks to construct complex ideas such as events and causal chains. This suggests that attempting to computationally identify long spans that represent such ideas may be problematic, particularly if human cognition itself has evolved to not rely on fixed span boundaries [11].

1.2.2 Why Not Break the Problem into Smaller Informational Chunks?

One possible strategy involves initially detecting smaller semantic chunks and then linking these chunks through relational structures to build more complex events or states. This approach parallels Event Extraction (EE) methodologies such as those used in JMEE [12] and PLMEE [13], in which event triggers are first identified and subsequently enriched by identifying event arguments. After such extraction, inter-event causal relationships can be classified. However, employing this approach requires specifically and intricately annotated datasets, which were not available for this project.

1.2.3 Why Not Approach Causal Extraction as an Extension of NER-RE?

This project explicitly aimed to simplify the causal extraction task into fundamental machine learning terms—specifically, identifying two spans of text, regardless of their length, which together form a causal relationship. Conceptually, this resembles traditional NER-RE tasks, which perform well on short-span extraction scenarios. However, identifying longer spans is considerably more difficult if one treats them as monolithic, clearly delineated structures. This difficulty arises because long spans rarely possess explicitly clear boundaries and instead tend to encode information contextually. Moreover, despite the simplicity of the desired annotation format, there currently exist no freely available causal datasets annotated with explicitly defined span boundaries suitable for this task.

1.3 Research Questions

This dissertation seeks to answer the following primary questions:

1. To what extent can existing discriminative NER-RE models be adapted effectively for long-span causal relation extraction?

2. What are the primary failure modes exhibited by discriminative IE models when applied to extracting long-span causal relations?
3. Which alternative modeling or methodological strategies show the greatest potential for overcoming identified limitations and should be prioritized for future research?

1.4 Research Outcome

This research effort resulted in the following findings:

- A discriminative joint span-based NER-RE model was successfully adapted for long-span causal relation extraction, though its performance was limited when applied to raw, unstructured natural language text.
- The model has less success with overly long or complex spans, implicit causality, longer range causality and complex causal chains. It performed best on shorter, explicit spans with direct, local causal relations.
- There is a clear lack of high-quality causal datasets, and existing resources vary widely in annotation style. This remains a significant barrier to progress and supports further research into weakly supervised annotation methods.
- Preliminary experimentation with large generative language models (LLMs) to simplify raw input into more explicit forms showed promise. These models were highly effective at summarising, translating, and rephrasing complex language into outputs such as simplified text, code, or JSON—especially when prompts included the original text to minimise hallucination. There appeared to be fewer errors when using generative models only for the front-end translation, leaving the structured identification and extraction to the discriminative model, rather than attempting end-to-end extraction via the LLM. However, this was only a preliminary exploration and not the core focus of the research.

2 Related Work

Causal Relation Extraction (CRE) is a more specific form of RE that focuses on identifying and classifying causal relationships between events or states. Unlike traditional RE, which primarily handles relationships between entities, CRE requires understanding the temporal and logical connections between events, which often involves deeper semantic reasoning. CRE overlaps with mainstream IE sub-tasks for NER-RE and EE, but has seen less research activity, potentially as this moves into other non-IE research areas such as causal reasoning that tend to processes knowledge graphs that have already been extracted previously.

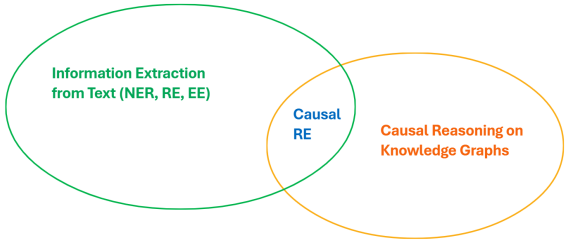


Figure 1: CRE Straddles the IE and Causal Reasoning Domains

CRE can be categorized into different types based on the explicitness and location of the causal relationship [14, 15, 16]:

- **Explicit Intra-sentential Causality:** The causal relationship is clearly expressed within a single sentence, often using connectives such as "because," "since," "therefore," or causative verbs like "cause," "lead to," or "result in."
- **Implicit Causality:** The causal relationship is implied but not explicitly stated, requiring inference based on contextual information and background knowledge. For example, "The rain stopped, and the sun came out."
- **Inter-sentential Causality:** The causal relationship spans multiple sentences, necessitating an understanding of connections between events across sentence boundaries. For example, "The stock market crashed. Investors lost a lot of money."

CRE presents several challenges compared to traditional RE:

- **Ambiguity:** Causal connectives and verbs can have multiple interpretations, making it difficult to determine if a relationship is truly causal.
- **Implicit Causality:** Many causal relationships are not explicitly stated, requiring models to infer connections from context and background knowledge.
- **Complex Event Form:** While states and many events take the noun/entity-proxy form, some events use the complex form with a trigger + arguments. This adds complexity to the causal relation identification task, eg. "The police locked up the criminal", requires either the whole event span to be detected and/or the components to be extracted.
- **Inter-sentential Dependencies:** Identifying causal relationships across sentence boundaries requires more complex systems to extract contextual signal important to an event pair. This becomes increasingly challenging over longer distances due to co-reference issues, context window sizes, as well as contextual signal dilution.
- **Lack of Annotated Data:** There is a general lack of quality annotated data for model training.

In this section, we offer a model summary table 2.1, followed by reviews of key IE models, organised by IE sub-task:

- Entity and Relation Extraction (NER-RE) 2.2
- Event Extraction (EE)2.3
- Multi-Task models which perform all three sub-tasks (NER-RE-EE) 2.4
- Causal Relation Extraction (CRE) 2.5

Lastly a review of strategies for Inter-Sentential RE focussed on transformer based models is offered 2.6.

Note that the foundational IE sub-tasks of NER, RE and EE are critical to the overall performance of Causal Relation Extraction (CRE), as CRE relies on accurately identifying entities, events and relationships. By examining the range of approaches applied to IE, we aim to highlight techniques that have proven effective in their respective areas, while also identifying those that could be adapted or re-purposed specifically for CRE systems.

2.1 Model Summary

The following table summarises some high performing recent models of interest on the various IE tasks along with their performance on associated datasets 4.2.1. Of note are are the "Model Type" and "Backbone" columns which refer to the general architecture and transformer backbone respectively. If available, links to the associated paper and source are given. Acronyms used in these columns are highlighted below:

- TR = Transformer (any type)
- DTR = Discriminative Transformer (pre-trained)
- GTR = Generative Transformer (pre-trained)
- PT = extra Pre-Training Required
- FT = Fine-Tuning Required
- Att = Attention Mechanism

Table 1: Models and Performance Across Datasets

Model	Model Type	Backbone	NER	RE	EE	CRE	Code	Paper
CMAN	**LSTM-Att**	**non-TR**	✓	✓	✗	✗	no code	paper
	F1-micro on "ADE"		89.4	81.1	-	-		
	F1-micro on "CoNLL04"		90.6	73	-	-		
DeepStruct	**GTR-PT-FT**	**GLM-10B**	✓	✓	✓	✗	code	paper
	F1-micro on "ACE05"		90	66.8	64.7	-		
	F1-micro on "ADE"		91.1	83.8	-	-		
	F1-micro on "CoNLL04"		90.7	78.3	-	-		

Continued on next page

Table 1 continued from previous page

Model	Model Type	Backbone	NER	RE	EE	CRE	Code	Paper
	F1-micro on "Genia2011"		80.8	-	-	-		
	F1-micro on "NYT"		95.9	93.3	-	-		
	F1-micro on "TACRED"		-	76.8	-	-		
DREEAM	**DTR-FT**	**RoBERTa**	✗	✓	✗	✗	code	paper
	F1-micro on "DocRED"		-	67.5	-	-		
DYGIE++	**DTR-FT-GNN**	**BERT**	✓	✓	✓	✗	code	paper
	F1-micro on "ACE05"		88.8	63.4	64.5	-		
	F1-micro on "Genia2011"		77.9	-	-	-		
	F1-micro on "SciERC"		67.5	48.4	-	-		
EXOBRAIN	**DTR-FT**	**RoBERTa**	✗	✓	✗	✗	no code	paper
	F1-micro on "Re-TACRED"		-	91.4	-	-		
	F1-micro on "TACRED"		-	75	-	-		
FourIE	**DTR-FT-GCN**	**BERT**	✓	✓	✓	✗	no code	paper
	F1-micro on "ACE05"		88.9	68.9	68.1	-		
GoLLIE	**GTR-FT**	**Code-LLaMA**	✓	✓	✓	✗	code	paper
	F1-micro on "ACE05"		89.6	70.1	70.3	-		
	F1-micro on "Onto. 5"		84.6	-	-	-		
GraphER	**DTR-FT-Att-GNN**	**ALBERT**	✓	✓	✗	✗	code	paper
	F1-micro on "ACE05"		89.8	68.4	-	-		
	F1-micro on "CoNLL04"		89.6	76.5	-	-		
	F1-micro on "SciERC"		69.2	50.6	-	-		
InstructUIE	**GTR-FT**	**FLAN-T5-11B**	✓	✓	✓	✗	code	paper
	F1-micro on "ACE05"		86.7	-	75	-		
	F1-micro on "ADE"		-	82.3	-	-		
	F1-micro on "CoNLL04"		-	78.5	-	-		
	F1-micro on "Genia2011"		74.7	-	-	-		
	F1-micro on "NYT"		-	90.5	-	-		
	F1-micro on "Onto. 5"		90.2	-	-	-		
	F1-micro on "SciERC"		-	45.1	-	-		
JMEE	**LSTM-GNN**	**non-TR**	✗	✗	✓	✗	no code	paper
	F1-micro on "ACE05"		-	-	69.6	-		
KEECI	**DTR-FT**	**BERT**	✗	✗	✗	✓	code	paper
	F1-micro on "SemEval2010 T8"		-	-	-	66		
MCDN	**Att-GRU-CNN**	**non-TR**	✗	✗	✗	✓	code	paper
	F1-micro on "AltLex"		-	-	-	82.5		
MCNN	**CNN**	**non-TR**	✗	✗	✗	✓	no code	paper
PFN	**PFN**	**non-TR**	✓	✓	✗	✗	code	paper
	F1-micro on "ACE05"		89	66.8	-	-		
	F1-micro on "ADE"		91.5	83.9	-	-		
	F1-micro on "NYT"		95.8	92.4	-	-		
	F1-micro on "SciERC"		66.8	38.4	-	-		
	F1-micro on "WebNLG"		98	93.6	-	-		
PLmarker	**DTR-FT**	**ALBERT**	✓	✓	✗	✗	code	paper

Continued on next page

Table 1 continued from previous page

Model	Model Type	Backbone	NER	RE	EE	CRE	Code	Paper
	F1-micro on "ACE05"		91.1	73	-	-		
	F1-micro on "Onto. 5"		91.9	-	-	-		
	F1-micro on "SciERC"		69.9	53.2	-	-		
PLMEE	**DTR-FT**	**BERT**	✗	✗	✓	✗	no code	paper
	F1-micro on "ACE05"		-	-	72.1	-		
RAG4RE	**GTR-RAG**	**FLAN-T5-11B**	✗	✓	✗	✗	code	paper
	F1-micro on "Re-TACRED"		-	73.3	-	-		
	F1-micro on "SemEval2010 T8"		-	14.1	-	-		
	F1-micro on "TACRED"		-	86.6	-	-		
RCEE	**DTR-FT**	**BERT**	✗	✗	✓	✗	code	paper
	F1-micro on "ACE05"		-	-	69.3	-		
REBEL	**GTR-FT**	**BART**	✗	✓	✗	✗	code	paper
	F1-micro on "ADE"		-	82.2	-	-		
	F1-micro on "CoNLL04"		-	75.4	-	-		
	F1-micro on "DocRED"		-	47.1	-	-		
	F1-micro on "NYT"		-	92	-	-		
	F1-micro on "Re-TACRED"		-	90.4	-	-		
SP	**DTR-FT**	**BERT**	✗	✓	✗	✗	no code	paper
	F1-micro on "SemEval2010 T8"		-	91.9	-	-		
	F1-micro on "TACRED"		-	74.8	-	-		
SPERT	**DTR-FT**	**BERT**	✓	✓	✗	✗	code	paper
	F1-micro on "ADE"		89.2	79.2	-	-		
	F1-micro on "CoNLL04"		88.9	71.5	-	-		
	F1-micro on "SciERC"		70.3	50.8	-	-		
SPLSTM	**LSTM**	**non-TR**	✗	✗	✗	✓	code	paper
	F1-micro on "AltLex"		-	-	-	82		
UIE	**GTR-PT-FT**	**T5-Large**	✓	✓	✓	✗	code	paper
	F1-micro on "ACE05"		85.8	66.1	64.1	-		
	F1-micro on "CoNLL04"		-	75	-	-		
	F1-micro on "SciERC"		-	36.5	-	-		
UniRel	**DTR-FT**	**BERT**	✗	✓	✗	✗	code	paper
	F1-micro on "NYT"		-	93.7	-	-		
	F1-micro on "WebNLG"		-	94.7	-	-		
USM	**DTR-FT**	**RoBERTa**	✓	✓	✓	✗	no code	paper
	F1-micro on "ACE05"		87.1	67.9	64.1	-		
	F1-micro on "CASIE"		-	-	53.6	-		
	F1-micro on "CoNLL04"		-	78.8	-	-		
	F1-micro on "NYT"		-	94.1	-	-		
	F1-micro on "SciERC"		-	37.4	-	-		

2.2 NER-RE Models

These models are focused two primary IE sub-tasks, NER (a span classification task) and RE (an span-pair relation classification task). Where the RE is for the relation between entities.

This area of IE has more research and code resources than other areas.

2.2.1 CMAN

(2018-[17, 18]) CMAN is a joint NER-RE classification model that uses non-transformer embeddings with Bi-LSTMs, self-attention, and cross-attention blocks, followed by a Conditional Random Field (CRF) head for NER and a multi-relation classification head for relation extraction (RE). Notable aspects of this model include the use of self-attention to enrich input word representations, cross-attention to incorporate NER label information into token representations, and a multi-relation classifier head structure.

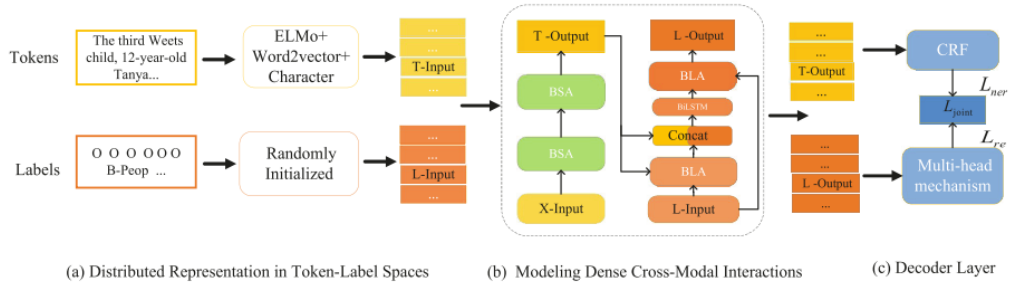


Figure 2: CMAN [17]

2.2.2 SpERT

(2020-[19]) SpERT (Span-based Entity and Relation Transformer) is a span-based pipeline model that performs joint entity and relation extraction using a BERT encoder. It predicts entities by classifying all spans within a sentence up to a fixed length. From the set of positively predicted spans, candidate span pairs are formed and classified for relations. While trained jointly with teacher-forced gold spans, inference operates in a pipe-lined fashion: relation extraction is only performed over predicted entity spans. This means relation prediction can suffer from missed entities, as relation candidates are never formed from spans not labeled as entities. Despite this limitation, SpERT achieves strong results on benchmarks like CoNLL04 through dense span enumeration and careful negative sampling.

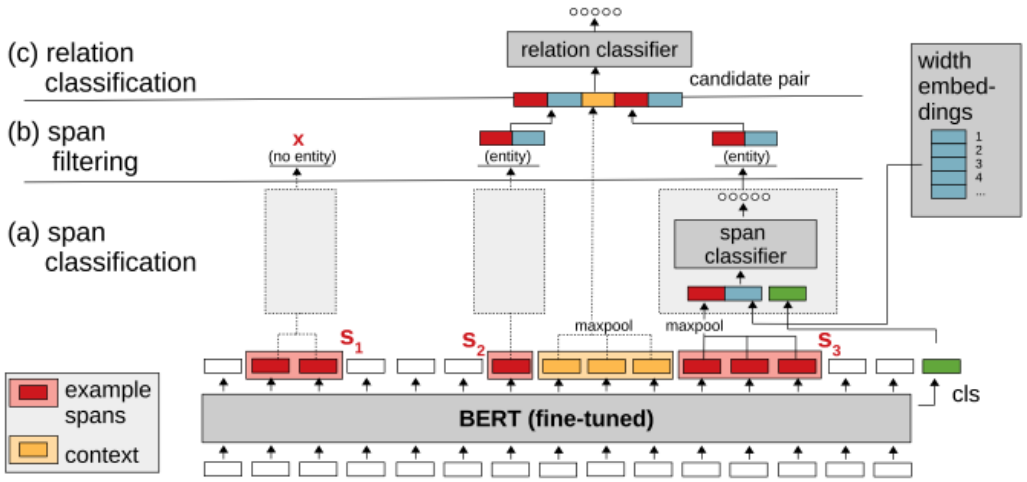


Figure 3: SpERT [19]

2.2.3 GraphER

(2024-[20, 21, 22]) GraphER is a joint NER-RE model. This model leverages transformer encodings and uses heuristics to form an initial noisy graph of entities and their relationships. It applies a graph structure-aware full attention mechanism (TokenGT [21]) to refine the initial graph representations before applying graph structure learning (GSL [22]) to prune and optimize the graph. This approach is designed to overcome two significant challenges seen in traditional GCN/GAT-based models: the bottleneck problem and the lack of structure learning. While this model is designed for NER-RE, it could be adapted for EE-CRE.

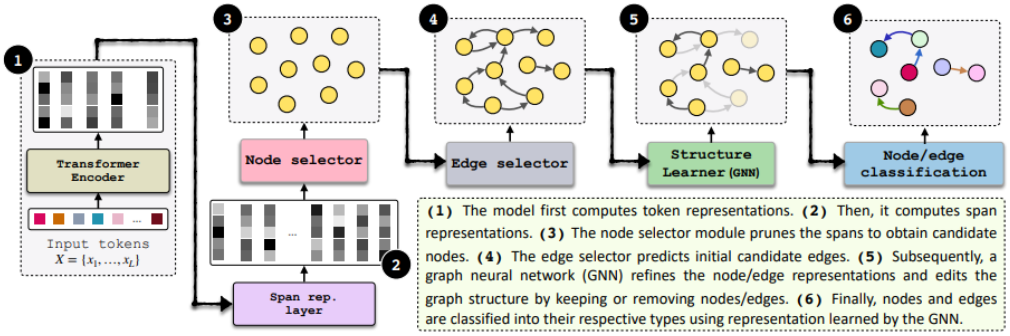


Figure 4: Knowledge Enhanced Event Causality Identification [20]

2.2.4 PL-marker

(2022-[23]) PLmarker is an LM based pipeline NER-RE span model based on efficient pre-marking entity and relation spans in input text prior to two independent transformer encoders,

inspired by [24]. Notable aspects of this approach are that it is the current SOTA for NER-RE, it's simplicity and the counter-intuitive fact that pre-encoder marking in a pipelined manner exceeds the performance of other more complex joint models.

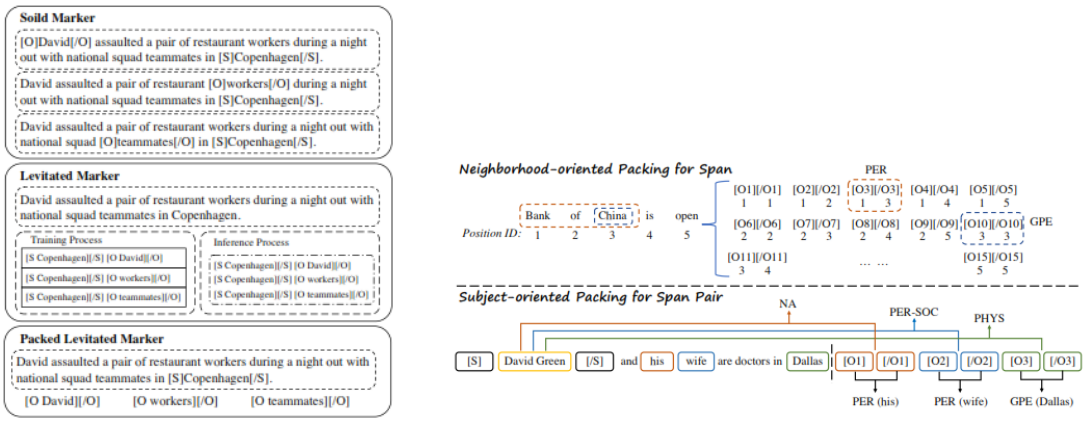


Figure 5: PLmarker [23]

2.3 EE Models

These models are focused the EE subtask. EE is typically broken down into Event Trigger identification/extraction/classification (a span classification task) and Event Argument identification/extraction and role classification (a span-pair relation classification task).

2.3.1 DEEB-RNN

(2018-[25]) DEEB-RNN, is a non-transformer word-token classification model that only performs Event Trigger detection and classification. It also requires NER results to work. While older, a notable aspect is the use of hierarchical attention to generate cross sentence (entire document) level representations, which are then combined with word and entity type embeddings prior to a Bi-GRU mixing and classification.

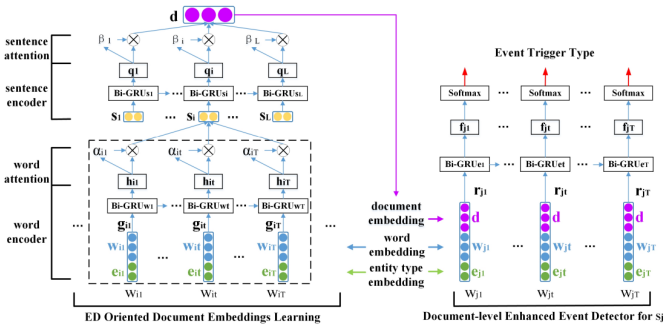


Figure 6: DEEB-RNN [25]

2.3.2 JMEE

(2018-[12]) JMEE is an Event Extraction model. This model utilized non-LM embeddings along with Bi-LSTM, GCN and self attention networks for joint event trigger and argument extraction. Note that it requires that the entities have already been extracted.

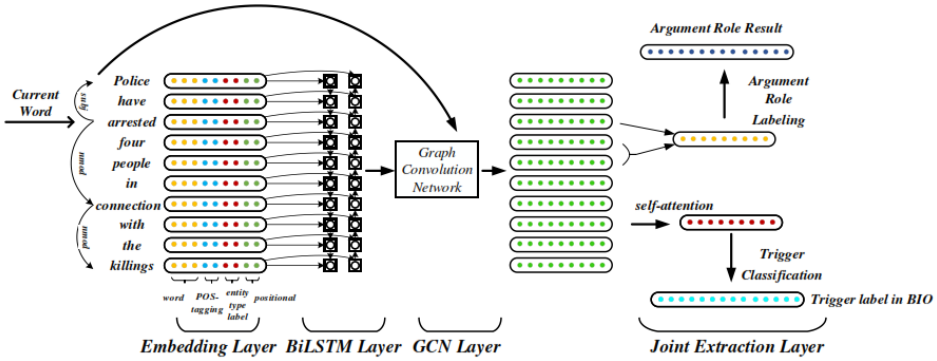


Figure 7: JMEE [12]

2.3.3 PLMEE

(2019-[13]) PLMEE is a model for Event Extraction, it operates as an pipeline model based on token classification with two independent BERT instances, one for the event trigger and the other for the arguments.

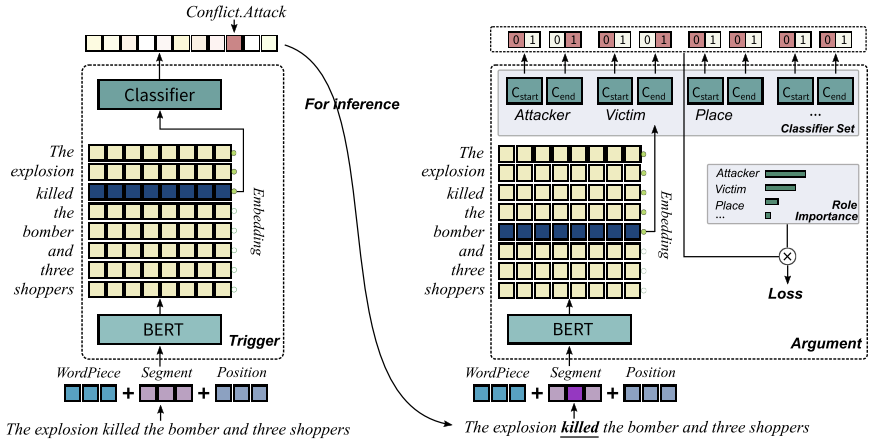


Figure 8: PLMEE [13]

2.4 Multi-Task Models

These models attempt to perform all three IE subtasks (NER, RE and EE).

2.4.1 DYGIE++

(2019-[26, 27]) DYGIE++ is a model for NER-RE and EE. It is a multi-task span-graph based joint model. It operates by processing transformer representations into span representations, then enriching them via graph techniques, allowing for the simultaneous extraction of entities, relations, and events. Another notable aspect is the use of overlapping input sequences to capture inter-sentence context.

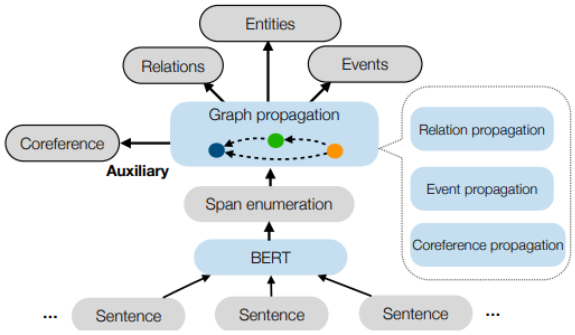


Figure 9: DYGIE++ [26]

2.4.2 InstructUIE

(2023-[28]) IntructUIE (Multi-task Instruction Tuning for Unified Information Extraction) is a high performing natural language prompting UIE Generative Transformer that can use fine-tuning to train just one model for multiple tasks. Additionally, it can be used in Few-shot mode with a few examples and no fine-tuning. The fundamental premise is to engineer descriptive natural language prompts to clearly outline each of the various IE tasks, each prompt has the following schema:

- **Instruction Clause:** detailing what task to perform, this details what IE task to perform, what elements to output and in what format along with any other specifics.
- **Option Clause:** detailing the result schema, i.e. the range of values that are allowed in the output.
- **Input Clause:** the text to process.
- **Output Clause:** has the labels (training) or blank to be filled by the model (inference).

Model fine-tuning (called instruction-tuning in the paper) is performed via labeled data for the main IE tasks (NER, RE, EE). Additionally these main tasks are broken down in to smaller auxiliary sub-tasks and added to the training to further help the model learn how to perform the more complex main tasks.

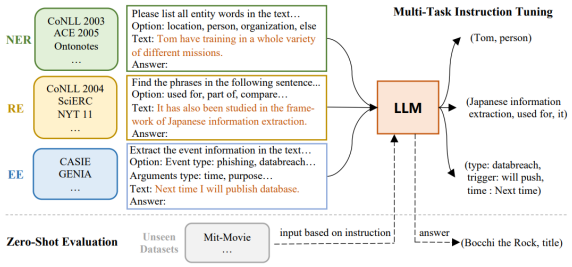


Figure 10: InstructUIE [28]

2.4.3 GoLLIE

(2024-[10]) GoLLIE (Guideline following Large Language Model for IE) is an code prompting UIE Generative Transformer. GoLLIE thus is built on a code optimized version of Llama which is better suited to parsing and generating structured outputs. Notable aspects of this model are:

- The model allows for fine-tuning over a range of datasets and tasks.
- They employ several regularization techniques to minimise hallucinations and make the model more robust and generalizable.
- The idea of using a code-in/code-out approach with code-specialized models is novel and makes a lot of sense for structured information extraction as well as specifying complex schema.

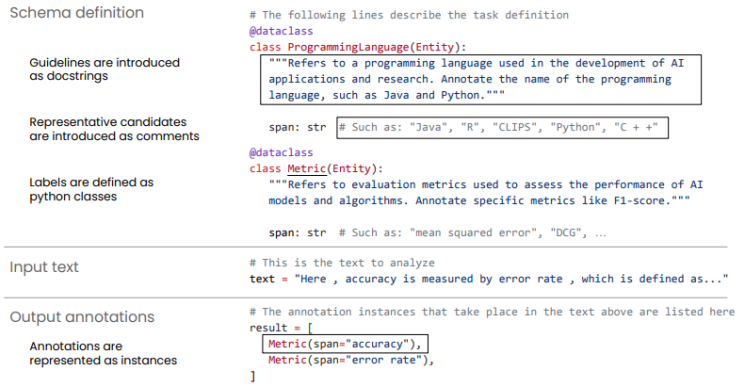


Figure 11: GoLLIE [10]

2.4.4 UIE

(2022-[9]) UIE (Unified Information Extraction) is a natural language prompting UIE Generative Transformer model. They frame the model as text-to-structure where all tasks are basically transformations of the input prompt text to a structured output text. They define an

input guideline format SSI (Structural Schema Instructor) and output format SEL (Structured Extraction Language), which are both pseudo JSON-like formats. The prompt contains the SSI along with the input text and the model generates the output in SSI format. A fundamental premise is that all IE tasks can be broken down to two granular components:

- **Spotting:** these are span semantic types to detect and extract.
- **Associating:** these are span relation semantic role types to detect and extract.

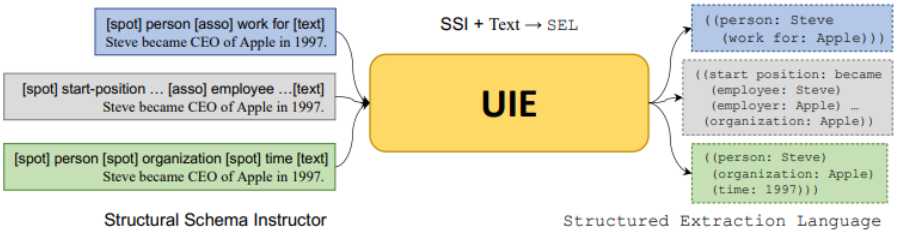


Figure 12: UIE [9]

2.5 CRE Models

These models are only concerned with identifying causal relations between spans of text. In all the papers reviewed the model require that the spans have already been identified.

2.5.1 MCNN

(2017-[29]) MCNN is an CNN-based non-transformer model for event causality classification. Notable features are how it combines external knowledge and contextual information with the event pair representations. Note that it requires that the event spans have already been extracted and decomposed to function.

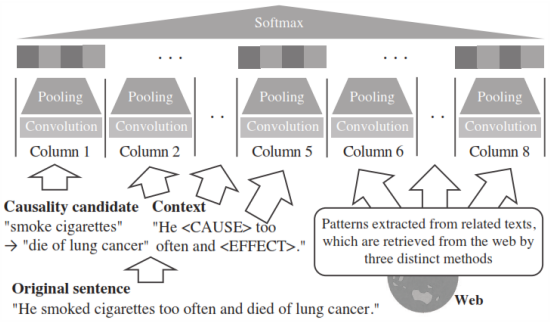


Figure 13: MCNN [29]

2.5.2 Stated Pair LSTM

(2017-[30]) The Stated Pair LSTM model leverages GloVe embeddings to process textual data along with event spans for causality classification. The model splits a given sentence into two segments, which are padded to equal lengths: (1) the first event span and (2) the concatenation of any relation trigger, a separator token and the second event span. These segments are each processed by separate LSTM networks. Noteworthy aspects of this model include:

- Its utilization of complete event spans without decomposing the events into smaller components, which simplifies the input structure.
- The model's adaptability to both explicit and implicit relational triggers, enabling it to handle varied linguistic contexts effectively.

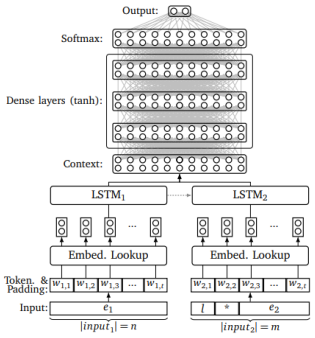


Figure 14: Stated Pair [30]

2.5.3 MCDN

(2021-[31]) uses non-transformer embeddings, an MHA layer, Multi-column CNNs to classify the causal relationship between two events. The MCDN model operates on the premise that two events and a relation trigger (denoted as 'AltLex') within the text are pre-identified. This identification allows the model to segment the text into three parts: before the relation trigger (BL), the relation trigger itself (L), and after the relation trigger (AL). Notable aspects of this model:

- It operates on event spans as opposed to decomposed event details
- While it needs an relation trigger, it could easily be adapted for implicit scenarios
- The way it processes the segment representations and combines them with the word level representations

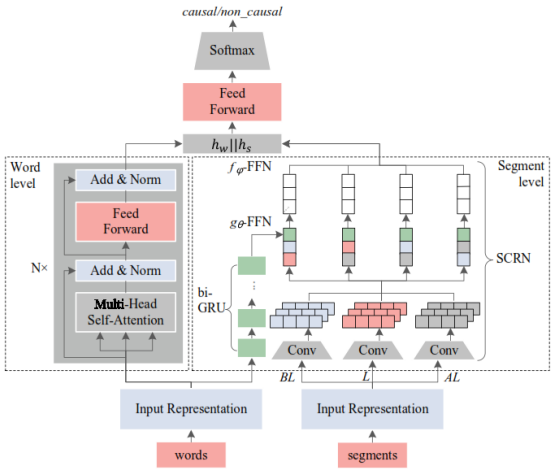


Figure 15: MCDN [31]

2.5.4 Knowledge Enhanced Event Causality Identification

(2020-[32]) This is an approach using an external knowledge base along with transformer encoders to build rich representations for an event-pair and to correspondingly classify it's causality. Note that this model expects the events spans have already been identified, not necessarily the event components though. Notable aspects of this model are the merging of external knowledge from a formal source, the novel method for building event-pair representations and the idea for the weighted combination of context only representations vs knowledge based enhanced representations.

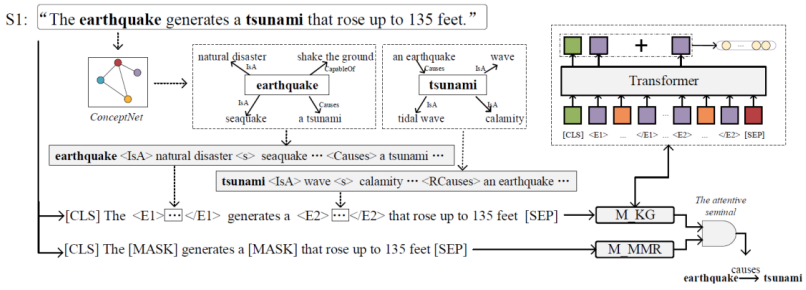


Figure 16: Knowledge Enhanced Event Causality Identification [32]

2.6 Inter-sentential RE

2.6.1 Discriminative Models for Inter-sentential RE

The specific issue of inter-sentential RE requires that the model can somehow encode contextual information over longer sequences of input text that a single sentence. Typically models work sentence by sentence, with transformer encoders like BERT having relatively small sequence length limits, which in-turn, limits the ability to include more inter-sentence

context. Recent advances have developed encoder transformers able to ingest longer sequences than BERT, such as Longformer [33] 4096 tokens and BigBird [34] 4096 tokens. With these limits the simple approach would be to just use a context window to include text before and after the current sentence as was done by [26, 23] and this may well be sufficient as the vast majority of direct inter-sentential events are going to be within a few sentences of each other.

However, for longer range relationships, there would need to be some kind of strategy to gather relevant contextual signal from long distances:

- [25] use a hierarchical attention mechanism to generate document embeddings for cross sentence context.
- [35] proposed a GCN approach, building a whole document graph with links between sequential sentence graphs. The idea of adding in the sentence to sentence relationships to create document level graph could potentially be investigated.
- A more recent model (SENDIR) [36] uses LSTMs in addtion to BERT. SENDIR determines event representations with inter-sentential context via a combination intra-sentence BERT token representations and inter-sentence Bi-LSTM token representations. NOTE: the Bi-LSTM is used as it doesn't suffer the BERT sequence length limits. The Bi-LSTM takes in the BERT output token representations for all sentences in the document, as such, it adds inter-sentential contextual information to the token representations.

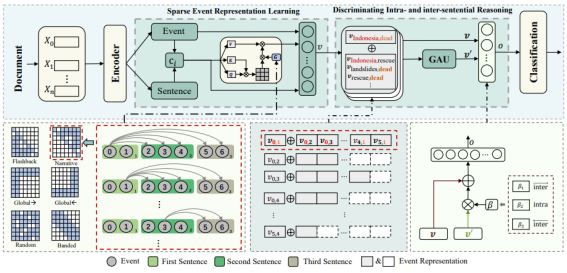


Figure 17: SENDIR [36]

2.6.2 Generative Transformers for Inter-sentential RE

Generative Transformers suffer from context window constraints, with the added issue for decoder only models that both the input and output tokens contribute to this context limit. This limit is increasing with some of the newer models [37], which potentially make it feasible to single prompt a model with an entire multi page report of text along with a comprehensive prompt including abundant examples and schema and still have ample context space for a long output.

However, the usefulness of the increased context window is currently limited due to the “lost in the middle problem” where the model seems to focus on the initial and last few 1000 tokens of a long document and gloss over the interior parts. This effect is well documented [38, 39] where they hypothesise that causes may range from training data bias, potentially issues with positional encoding, possible internal summarization or attention window techniques to manage speed and compute resources. Fig 19 shows the performance of GPT-4o for extraction of 20 distributed unique keys for various clusterings of the keys (100 meaning all in the same area, 0 meaning highly distributed) vs corpus length. This shows the performance degrades as the corpus length increases and the distribution of the keys becomes more random.

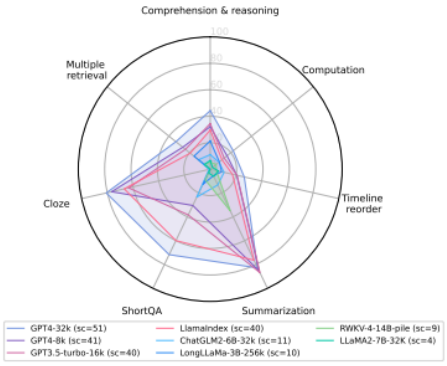


Figure 18: Generative Transformer Long Dependency Performance [38]

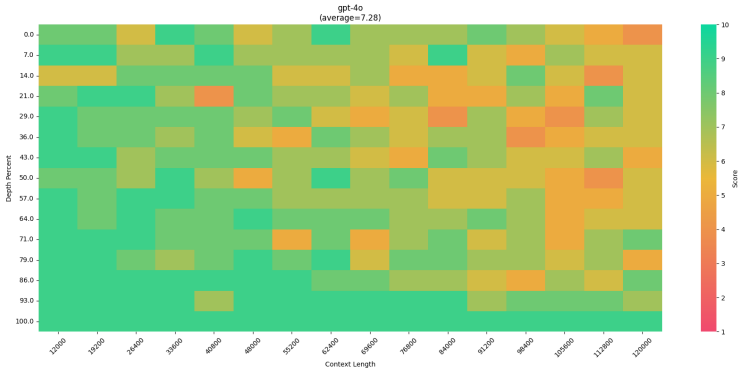


Figure 19: 20 needles in the haystack performance for GPT4o [39]

3 Methodology

3.1 Model Architecture

This work explores how span and span-pair triplet extraction architectures, inspired by models such as SpERT [19] and GraphER [20], can be adapted for long-span causal relation extraction (CRE). The goal is to evaluate whether the structural simplicity of this class of model remains effective when applied to CRE's more complex span dynamics. As shown in Fig 20 the overall model is a pipeline structure.

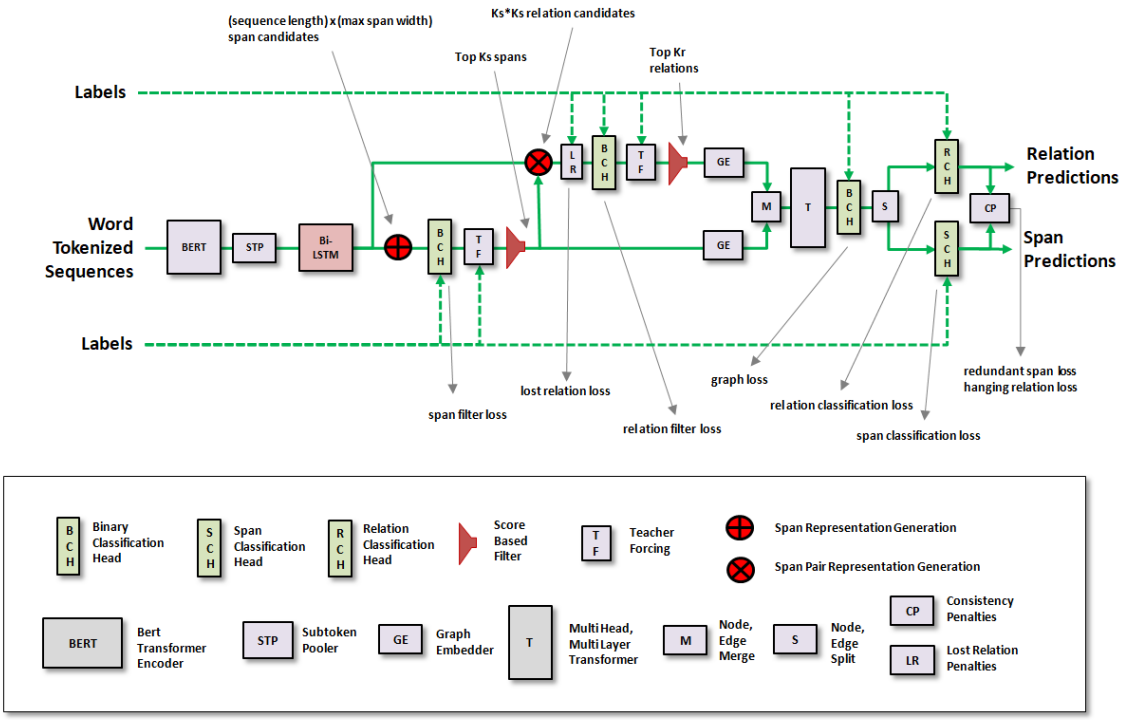


Figure 20: Model, Original Variant

The model contains the following components in order:

- **Pretrained Transformer Encoder Layer**
- **BiLSTM Layer**
- **Span Generator and Filter:**
  - *Either:* **Token Tagging Layer + Span Generator + Span Filter**
  - *Or:* **Span Generator + Binary Filter Head + Span Filter**
- **Relation Generator + Binary Filter Head + Relation Filter**
- **Graph Embedder + Graph Transformer**

· **Span Classification Head + Relation Classification Head**

These components are described in more details in the following sections.

3.1.1 Pretrained Transformer Encoder Layer (BERT

The model uses a pretrained BERT encoder [40] from HuggingFace Transformers [41] to encode input text at the subword level. Since BERT operates on subword tokenization (e.g., WordPiece), a conversion step is required to obtain word-level token representations. This is achieved via max-pooling over the hidden states of each word's subword pieces:

$$\mathbf{h}_i^{(\text{word})} = \max_{j \in \mathcal{S}(i)} \mathbf{h}_j^{(\text{sub})}$$

where $\mathcal{S}(i)$ is the set of subword token indices corresponding to word token i, and $\mathbf{h}_j^{(\text{sub})}$ is the hidden state from BERT for subword token j.

This pooling strategy was selected to simplify downstream processing and eliminate the need to adapt all downstream modules to subword boundaries. Other pooling methods, such as first-subword or first-last averaging (as implemented in Flair [42]), were also considered, but empirical results showed that max-pooling performed best in this context. Similar pooling techniques have been used successfully in prior work such as [20, 43].

3.1.2 BiLSTM Layer

An optional BiLSTM layer [44] is applied on top of the word token embeddings to further enrich their contextual representation. This design choice was inspired by its successful use in relation extraction pipelines such as [20, 45].

Given a sequence of word-level token embeddings $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$, the BiLSTM outputs:

$$\mathbf{h}_i^{(\text{BiLSTM})} = [\overrightarrow{\mathbf{h}}_i;\ \overleftarrow{\mathbf{h}}_i]$$

where $\overrightarrow{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ are the hidden states from the forward and backward LSTM respectively at position i, and $[\cdot;\cdot]$ denotes vector concatenation.

The output of the BiLSTM is further processed via layer normalization and dropout. A residual connection is optionally applied between the input and output of the BiLSTM layer.

3.1.3 Token Tagger Layer + Filter

This layer provides an alternative to brute-force span search by predicting candidate spans directly from token-level boundary labels. It is significantly more efficient for longer sequences and offers comparable performance. The BIO tagging scheme was considered to be too restrictive, hence two custom tagging strategies were implemented, both designed to predict span boundaries and deal with overlaps:

- **BE Multiclass Token Tagging (Begin-End):** A binary vector of length 2 is predicted for each token t_i, representing `[Begin, End]` probabilities:

$$\hat{\mathbf{y}}_i = [\hat{y}_i^{(B)},\ \hat{y}_i^{(E)}] = \sigma(W \cdot h_i + b)$$

  All valid pairs (i, j) such that $\hat{y}_i^{(B)} > 0$, $\hat{y}_j^{(E)} > 0$, and $j > i$, with span width $\leq w_{\text{max}}$, are used to construct candidate spans. The span filter score is:

$$\text{score}_{i,j} = \frac{1}{2}\left(\hat{y}_i^{(B)} + \hat{y}_j^{(E)}\right)$$

- **BECO Uniclass Token Tagging (Begin-End-Combined-Other):** Each token is assigned one of 4 mutually exclusive classes using softmax:

$$\hat{y}_i = \text{softmax}(W \cdot h_i + b), \quad \hat{y}_i \in \{\text{B}, \text{E}, \text{C}, \text{O}\}$$

  Candidate spans are constructed by pairing predicted B and E tags, or directly using C (Combined) tags as single-token spans. Valid span conditions remain the same.

Finally, the top-K candidate spans, ranked by their filter scores, are selected for downstream processing.

The token-level classification loss is:

- **Binary Cross Entropy (BCE)** for BE tagging:

$$\mathcal{L}_{\text{BCE}} = -\sum_i \left[y_i^{(B)} \log \hat{y}_i^{(B)} + y_i^{(E)} \log \hat{y}_i^{(E)}\right]$$

- **Cross Entropy (CE)** for BECO tagging:

$$\mathcal{L}_{\text{CE}} = -\sum_i \log \hat{y}_i^{(c_i)} \quad \text{where } c_i \in \{\text{B}, \text{E}, \text{C}, \text{O}\}$$

Note: Teacher forcing was optionally applied during training, ensuring that gold spans were retained regardless of filter score, without affecting loss computation.

3.1.4 Span Representation Generator

Spans are of variable token lengths, but for modeling, span representations must be converted into fixed-size hidden vectors. Span representations are constructed from candidate spans—either the top-K spans selected by a token tagger, or all spans (up to a width limit) if no tagger is used. Several construction strategies were implemented:

- **First-Last:** Concatenate the word token representations for the start and end of the span, then re-project to the hidden dimension:

$$\text{span\_rep} = \text{FFN}\left([h_{\text{start}};\ h_{\text{end}}]\right)$$

- **Maxpool-CLS-Width:** Concatenate the max-pooled span representation, the BERT `[CLS]` embedding, and a learnable span width embedding:

$$\text{span\_rep} = \text{FFN}\left([\text{maxpool}(h_{\text{span}});\ h_{[\text{CLS}]};\ \text{width\_emb}]\right)$$

- **First-MaxpoolBetween-Last-CLS-Width:** Concatenate the pooled start, end, and inner span segments, plus `[CLS]` and width embeddings:

$$\text{span\_rep} = \text{FFN}\left([\text{pool}_{\text{start}};\ \text{pool}_{\text{end}};\ \text{pool}_{\text{inner}};\ \text{width\_emb};\ h_{[\text{CLS}]}\right)$$

  Where:

  - $\text{pool}_{\text{start}}$ is max-pooled over the first w tokens of the span, $w = \min(1, \alpha \cdot \text{span\_width})$
  - pool_{end} is max-pooled over the last w tokens of the span
  - $\text{pool}_{\text{inner}}$ is max-pooled over remaining span tokens (or $\text{pool}_{\text{start}}$ if none)

- **AttentionPooling-CLS-Width:** Apply self-attention pooling over the span tokens, then concatenate with `[CLS]` and width embedding:

$$\text{span\_rep} = \text{FFN}\left([\text{attnpool}(h_{\text{span}});\ h_{[\text{CLS}]};\ \text{width\_emb}]\right)$$

In all cases, $\text{FFN}(\cdot)$ is a two-layer feedforward network with an intermediate expansion, ReLU, dropout, and final projection to the model hidden dimension.

3.1.5 Binary Span Filter Layer

This layer applies a configurable binary classification head to each span representation to produce a filtering score. Two configurations were supported:

- **Single-logit mode:** A single scalar logit is produced for each span representation $\mathbf{SR}_i$:

$$s_i = \mathbf{w}^\top \mathbf{SR}_i + b$$

  where s_i is used directly as the filtering score. The label $y_i \in \{0, 1\}$ is the binarised gold label for the span (1 = keep, 0 = discard). The loss is Binary Cross Entropy (BCE):

$$\mathcal{L}_{\text{BCE}} = -\left[y_i \cdot \log \sigma(s_i) + (1 - y_i) \cdot \log(1 - \sigma(s_i))\right]$$

- **Double-logit mode:** A 2-class classifier outputs two logits for each span:

$$\mathbf{s}_i = W^\top \mathbf{SR}_i + \mathbf{b} \quad \text{where} \quad \mathbf{s}_i \in \mathbb{R}^2$$

  The filtering score is computed as the logit delta:

$$\text{score}_i = s_i^{(1)} - s_i^{(0)}$$

  The binarised label $y_i \in \{0, 1\}$ is used with Cross Entropy (CE) loss:

$$\mathcal{L}_{\text{CE}} = -\log\left(\frac{e^{s_i^{(y_i)}}}{e^{s_i^{(0)}} + e^{s_i^{(1)}}}\right)$$

In both cases, the top-K scoring spans are selected for downstream processing. If a token-level tagger is used prior to this layer, the top-K here may be set lower—but this setup did not show consistent benefit and was typically not used in combination.

Note: Teacher forcing was optionally applied during training, in which case gold spans were always passed through the filter without affecting the loss.

3.1.6 Relation Representation Generator

After the top K spans are generated, all possible directed relations are formed via the Cartesian product of the span set, producing K^2 candidate relations. This stage introduces a quadratic scaling bottleneck relative to the span count.

Each candidate relation is represented using a combination of its head and tail span representations and optional context. The following configurations were implemented:

- **No Context:** Concatenate the head and tail span representations, then re-project to the hidden dimension in the hope that all required signal is contained within the head and tail:

$$\text{rel\_rep} = \text{FFN}([\mathbf{SR}_h;\ \mathbf{SR}_t])$$

- **Between Context:** Concatenate the head and tail span representations with a pooled representation of the tokens between the two spans. This is commonly used for NER-RE models:

$$\text{rel\_rep} = \text{FFN}([\mathbf{SR}_h;\ \mathbf{SR}_t;\ \text{pool}(T_{\text{between}})])$$

- **Window Context:** Concatenate the head and tail span representations with pooled representations from windows before and after each, this was an adaptation to account for longer spans and more varied context:

$$\text{rel\_rep} = \text{FFN}([\mathbf{SR}_h;\ \mathbf{SR}_t;\ \text{pool}(T_{\text{pre}});\ \text{pool}(T_{\text{post}})])$$

- **Between + Window Context:** Combine both the between-span and windowed context tokens:

$$\text{rel\_rep} = \text{FFN}([\mathbf{SR}_h;\ \mathbf{SR}_t;\ \text{pool}(T_{\text{between}});\ \text{pool}(T_{\text{pre}});\ \text{pool}(T_{\text{post}})])$$

Note: The function $\text{pool}(\cdot)$ refers to a configurable pooling mechanism applied over a token set. Variants tested included max pooling, self-attention pooling, and cross-attention pooling with the head or tail span representations as queries.

3.1.7 Binary Relation Filter Layer

This layer applies a configurable binary classification head to each relation representation to produce a filtering score. Two configurations were supported:

- **Single-logit mode:** A single scalar logit is produced for each relation representation $\mathbf{RR}_i$:

$$s_i = \mathbf{w}^\top \mathbf{RR}_i + b$$

where s_i is used directly as the filter score. The binarized label $y_i \in \{0, 1\}$ indicates whether the relation is positive (1) or negative (0). The loss is Binary Cross Entropy (BCE):

$$\mathcal{L}_{\text{BCE}} = -\left[y_i \cdot \log \sigma(s_i) + (1 - y_i) \cdot \log(1 - \sigma(s_i))\right]$$

- **Double-logit mode:** A 2-class classifier outputs two logits:

$$\mathbf{s}_i = W^\top \mathbf{RR}_i + \mathbf{b} \quad \text{where} \quad \mathbf{s}_i \in \mathbb{R}^2$$

The filter score is calculated as the logit difference:

$$\text{score}_i = s_i^{(1)} - s_i^{(0)}$$

With the binarized label $y_i \in \{0, 1\}$, the classification loss is Cross Entropy:

$$\mathcal{L}_{\text{CE}} = -\log\left(\frac{e^{s_i^{(y_i)}}}{e^{s_i^{(0)}} + e^{s_i^{(1)}}}\right)$$

In both modes, the top-K relations (by score) are retained for downstream processing.

Note: Teacher forcing was optionally applied during training, ensuring that gold relations bypassed filtering without contributing to the loss.

3.1.8 Graph Embedder

The graph module is optional. When enabled, the top-K span representations and top-K relation representations are used to form the initial graph structure. Each span representation $\mathbf{SR}_i$ is stamped with a learned span node embedding $\mathbf{e}_{\text{span}}$, and each relation representation $\mathbf{RR}_j$ is stamped with a learned edge embedding $\mathbf{e}_{\text{rel}}$:

$$\tilde{\mathbf{SR}}_i = \mathbf{SR}_i + \mathbf{e}_{\text{span}}, \quad \tilde{\mathbf{RR}}_j = \mathbf{RR}_j + \mathbf{e}_{\text{rel}}$$

This addition allows the transformer to differentiate span nodes from relation edges. This typed graph embedding approach is adapted from [20].

3.1.9 Graph Transformer

The stamped span and relation representations are stacked along the graph axis (i.e., treated as distinct graph elements in a flat sequence) to form the input to the graph transformer:

$$\mathbf{G} = \text{concat}(\tilde{\mathbf{SR}}_{1:K_s},\ \tilde{\mathbf{RR}}_{1:K_r}) \in \mathbb{R}^{(K_s+K_r)\times d}$$

where d is the hidden dimension, K_s is the number of span candidates, and K_r is the number of relation candidates.

This combined representation is passed through a multi-layer transformer encoder with multi-head self-attention (MHA):

$$\mathbf{G}' = \text{Transformer}(\mathbf{G})$$

3.1.10 Graph Binary Filter Layer

This layer applies a configurable binary classification head to each graph element (node, edge) representation to produce a filtering score. The filtering score is not used to prune the graph further, it is used only to calculate an additional loss vs the labels, called the graph loss.

Finally, the graph representations $\mathbf{G}'$ are split back into updated node (span) and edge (relation) representations:

$$[\mathbf{SR}'_1, \ldots, \mathbf{SR}'_{K_s}],\ [\mathbf{RR}'_1, \ldots, \mathbf{RR}'_{K_r}] = \text{split}(\mathbf{G}')$$

3.1.11 Span Classification Head

Each node (span) representation is passed through a linear classification head to produce logits over N_s span types:

$$\mathbf{z}_i^{\text{span}} = W_{\text{span}} \cdot \mathbf{SR}'_i + b_{\text{span}} \in \mathbb{R}^{N_s}$$

A Cross Entropy loss is applied, assuming each span belongs to exactly one class (uniclass setup):

$$\mathcal{L}_{\text{span}} = -\log\left(\frac{e^{z^{(y_i)}}}{\sum_{j=1}^{N_s} e^{z^{(j)}}}\right)$$

3.1.12 Relation Classification Head

Each edge (relation) representation is passed through a linear classification head to produce logits over N_r relation types:

$$\mathbf{z}_j^{\text{rel}} = W_{\text{rel}} \cdot \mathbf{RR}'_j + b_{\text{rel}} \in \mathbb{R}^{N_r}$$

As multiple relation types may apply to each relation (multilabel setup), a sigmoid activation is applied element-wise, and Binary Cross Entropy is used:

$$\mathcal{L}_{\text{rel}} = -\sum_{k=1}^{N_r} [y_k \log \sigma(z_k) + (1 - y_k) \log(1 - \sigma(z_k))]$$

3.2 Model Variants

Various model variants were coded and tested, these have been broken down into 2 main variants.

3.2.1 Original Variant

This is the original configuration 20 which uses brute force span searching in the initial span filtering stage. This works well for smaller span widths and sequence lengths. The resource limitation comes from the initial span representation generation algorithm. The number of possible spans to search being proportional to $sequence.length \times max.span.width$. This was loosely inspired by the model presented in [20] which essentially uses intermediate binary filter layers for both spans and relations as a form of smart negative sampling to shortlist the spans/relations to smaller sets based on the likelihood of the span/relation being significant. The graph transformer and graph filter layer are optional.

The thoughts behind these filter layers as described by [20] was to form an initial graph with spans being nodes and relations being edges. The graph transformer then being a final step to score the graph structure wholistically vs the labels. The pipeline for the original variant is thus:

- shared BERT pre-trained encoder with post encoder sub-token pooling followed by an optional LSTM layer
- span representation generator and optional random negative case sampler
- binary span filter layer which produces both a span filtering loss and a filtering score for each candidate span. This is used for smart filtering (negative sampling) down to the best Ks spans for downstream processing
- relation representation generator and optional random negative case sampler
- binary relation filter layer which produces both a relation filtering loss and a filtering score for each candidate relation. This is used for smart filtering (negative sampling) down to the best Kr relations for downstream processing.
- optional graph transformer and binary graph filter layer which produces a graph structure filtering loss vs the span/relation labels.
  No logit based score filtering/pruning is done in the graph stage.
- output span and relation classification heads

Sub variations of this model were:

- use a separate BERT instance for the span generator and relation generator.
- disabling of teacher forcing after a certain point in the training process and a lost relation loss to be enabled as a replacement.
- enabling and disabling different combinations of the lstm and graph layers

- use training span/relation negative sampling in addition to the binary filtering.
  NOTE: random negative sampling can't just replace the smart negative sampling as it is not used during inference mode, leading to the full size tensors of possible spans/relations being passed downstream, which can quickly cause memory issues. Smart negative sampling, however, trains a binary filter head which will be used during inference, thus reducing memory use explosion after the initial warm-up stage.

3.2.2 Long span Variant

This variation swapped out the binary span filter layer with the a token tagging layer, see Fig 21. This is primarily to deal with the explosion in memory usage as the sequence length and span widths get longer. There was minimal performance variation from the original variant quantitatively or qualitatively. The token tagging scheme used also allowed for overlapping spans. For longer sequences and span widths, token tagging (depending on the tagging scheme) would typically produce < 1000 span candidates where the brute force method for the same configuration would produce > 10000 candidates.

Ablations on the token tagger specifically revolved around the tagging scheme with two variants being tested (BE and BECO). Standard BIO was not reviewed as it did not fit the usage case.

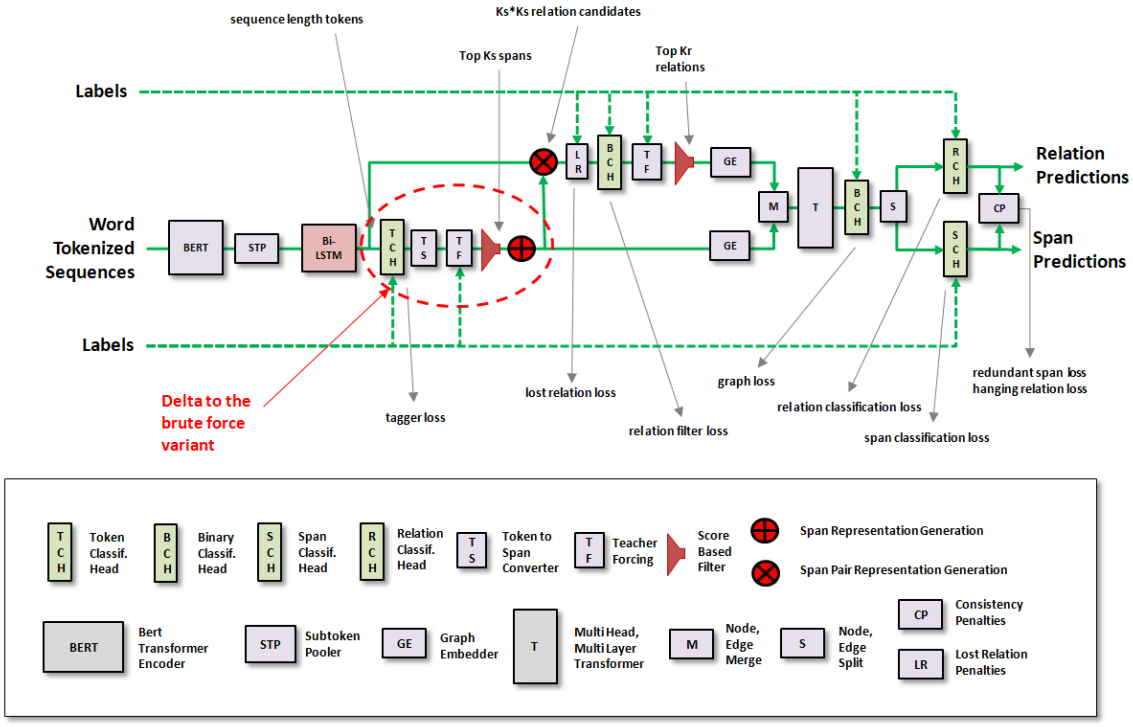


Figure 21: Model, Tagging Variant

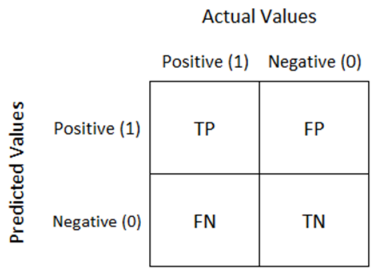
3.2.3 Other variants

Some other variants were trialed such as marking spans and relations. This involved marking the top-K spans and relations in the original sequence with tags and passing the modified sequence through an additional pre-trained encoder transformer for each span/relation. More specifically, the top K spans produced after the token tagging layer were marked in the original sequences and run through a separate BERT instance and the pooler token (CLS) embeddings collected as the span representations. These span representations were then used as inputs to the relation representation generator, which then passed through the binary relation filter layer to produce the top K relations. There again, each relation (pair of spans) was marked on the original sequence and passed through a separate BERT instance. The pooler (CLS) embeddings collected to form the new representation for each of the top K relations. The idea here was to test whether pre-marking spans/relations and leveraging BERT could improve the span/relation representations in any way as good relation F1 results were shown in [46]. However, it became apparent that the numbers in that particular paper were misleading as they effectively gave the model the label span-pairs for marking, thus only testing the ability of the marking scheme to determine causality given the labels, this is very different to the whole end to end pipeline as was being done with this model. As this was extremely resource inefficient, it was only tested as a proof of concept with results indicating that it offered no benefit to performance metrics in the full end to end pipeline. Aside from these modifications the pipeline and optional components remained similar to previous variants.

4 Experimental Setup

4.1 Evaluation Metrics

As IE is primarily a discriminative task by nature, model performance viewed via classification metrics. The most common classification metric for IE models is F1 as the positive outcome is the focus and the positive classes are typically very sparse. Additionally, for multi-class scenarios, F1-macro will give a more balanced view of the F1 score (as opposed to F1-micro) as it averages across all classes irrespective of sample counts. If it is imperative to account for the True Negatives, then MCC can be a more balanced metric.



Accuracy (Acc): $= \frac{TP+TN}{TP+TN+FP+FN}$

Recall (TPR): $= \frac{TP}{TP+FN}$

Precision: $= \frac{TP}{TP+FP}$

F1 Score: $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Matthews Correlation Coefficient (MCC): $= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$

Figure 22: Classification Metrics

4.2 Datasets

Datasets were a significant issue for this research as there simply do not exist a selection of free quality datasets for causal relation extraction annotated in a compatible way. This actually became one of the major stumbling blocks of the project as the model and the annotations it is designed for are tightly coupled. Some of the key points regarding datasets are:

- No compatible causal relation datasets were available. The closest found were curated by the Unicausal project [46], however each of the datasets from this project had their own issues including excessive annotation noise (spans too long, too short, irrational spans, including the causal phrases). These problems were severe enough to render each and every dataset effectively unusable. Several other known datasets were paid [47, 48, 49], so no analysis was attempted on these.
- Overlap with target domain. The target domain is geology reports which formal language with a very specific geology-centric vocabulary. No dataset had anything related to this kind of text. Attempts were made in the initial phase of the project to develop code to extract pdf text from geological reports, chunk it and prepare it for annotations. However time constraints limited the depth of this investigation and it was not developed further.
- Self-Annotation and Time Constraints: some minor self annotations were attempted to at least form a core sample dataset for causal data. A lightweight annotation tool was developed to enable rapid labeling under time constraints. A small test dataset was annotated for testing purposes. It is clear that with more time and resources, a higher quality and larger dataset could have been curated.

4.2.1 Datasets Overview

The following table details some of the most prevalent datasets, which models use it, what tasks it is used for and links to the relevant paper and source.

Table 2: Summary of Datasets

Dataset	Models	UC*	NER	RE	EE	CRE	Source	Paper	Cost
ACE05	DeepStruct, DYGIE++, FourIE, GoLLIE, GraphER, InstructUIE, JMEE, PFN, PLmarker, PLMEE, RCEE, UIE, USM	✗	✓	✓	✓	✗	source	paper	$3100

Continued on next page

Table 2 continued from previous page

Dataset	Models	UC*	NER	RE	EE	CRE	Source	Paper	Cost
ADE	CMAN, DeepStruct, InstructUIE, PFN, REBEL, SPERT	✗	✓	✓	✗	✓	source	paper	Free
AltLex	MCDN, SPLSTM	✓	✗	✗	✗	✓	source	paper	Free
Because 2	-	✓	✗	✗	✗	✓	source	paper	?Free?
CASIE	USM	✗	✓	✓	✓	✓	source	paper	Free
CausalTimeBank	-	✓	✗	✗	✓	✓	source	paper	Free
CoNLL04	CMAN, DeepStruct, GraphER, InstructUIE, REBEL, SPERT, UIE, USM	✗	✓	✓	✗	✗	source	paper	Free
DocRED	DREEAM, REBEL	✗	✓	✓	✗	✗	source	paper	Free
EventStoryLine	-	✓	✗	✗	✓	✓	source	paper	Free
Genia2011	DeepStruct, DYGIE++, InstructUIE	✗	✓	✗	✗	✗	source	paper	?Free?
NYT	DeepStruct, InstructUIE, PFN, REBEL, UniRel, USM	✗	✓	✓	✗	✗	source	paper	NA
Onto. 5	GoLLIE, InstructUIE, PLmarker	✗	✓	✗	✗	✗	source	paper	?Free?
PDTB	-	✓	✗	✗	✓	✓	source	paper	$760
Re-TACRED	EXOBRAIN, RAG4RE, REBEL	✗	✓	✓	✗	✗	source	paper	Free
SciERC	DYGIE++, GraphER, InstructUIE, PFN, PLmarker, SPERT, UIE, USM	✗	✓	✓	✗	✗	source	paper	Free
SemEval2010 T8	KEECI, RAG4RE, SP	✓	✓	✓	✗	✓	source	paper	Free
TACRED	DeepStruct, EXOBRAIN, RAG4RE, SP	✗	✓	✓	✗	✗	source	paper	35
WebNLG	PFN, UniRel	✗	✓	✓	✗	✗	source	paper	Free

4.2.2 Altlex [55]

Altlex has a single causal pair annotated per observation. The sequence is effectively cut in two parts with each part making up the spans. Thus many cases do not have any real rational behind the span boundaries, making it less than ideal for use with the model as these would be counted as positive span cases and cause confusion during model training. To be useful, this dataset would need re-annotation of the span boundaries. As can be observed from the POS breakdown, the spans are not concentrated on one part of speech.

*A recent (2023) resource in the realm of CRE specific datasets is UniCausal [46]. The authors have processed 6 causal datasets and released the 5 free ones on their github page with the goal of standardising the format of each of the component datasets. The 5 free datasets are: AltLex [50]; BECAUSE 2.0 [51]; CausalTimeBank [52]; EventStoryLine [53]; SemEval2010 T8 [54]. The single paid datset is PDTB [48].

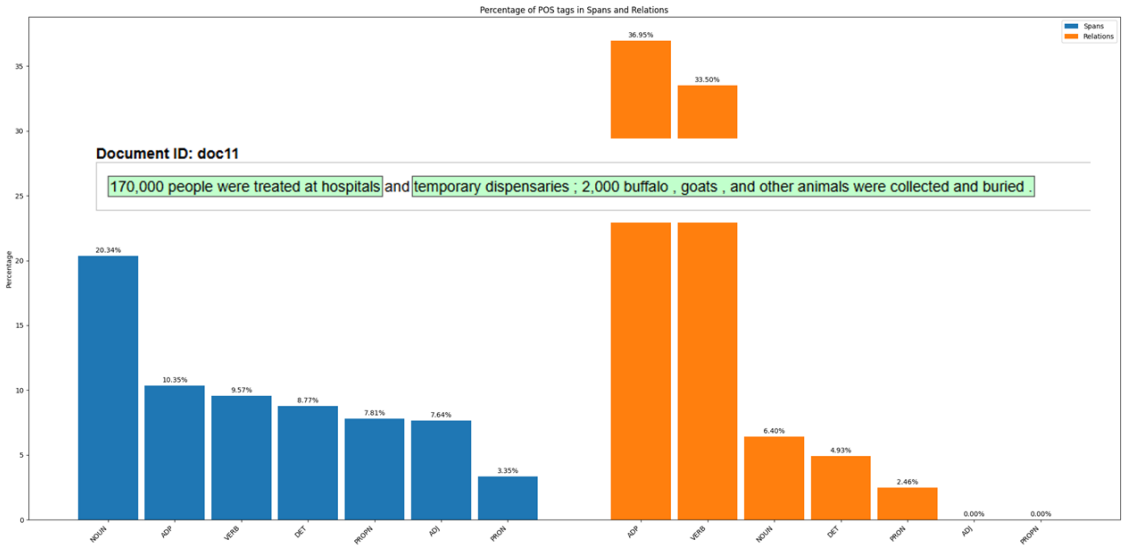


Figure 23: Altlex POS Breakdown

4.2.3 SemEval [54]

SemEval has many samples, but has very short spans (up to 4 words) annotated causal pairs. As can be seen from the POS breakdown, the annotated spans are almost all entities (see the following POS breakdown). This dataset has approximately 90% negative cases, where the negative cases are just unrelated and irrational spans, i.e. not forming an event/state, just randomly chosen entities. It is only for the few positive cases the annotated spans typically are the entity form of events or states or entity proxies for events and states. Additionally all causal pairs are explicitly linked. Thus this dataset again has limited usage for the model and would need significant revision.

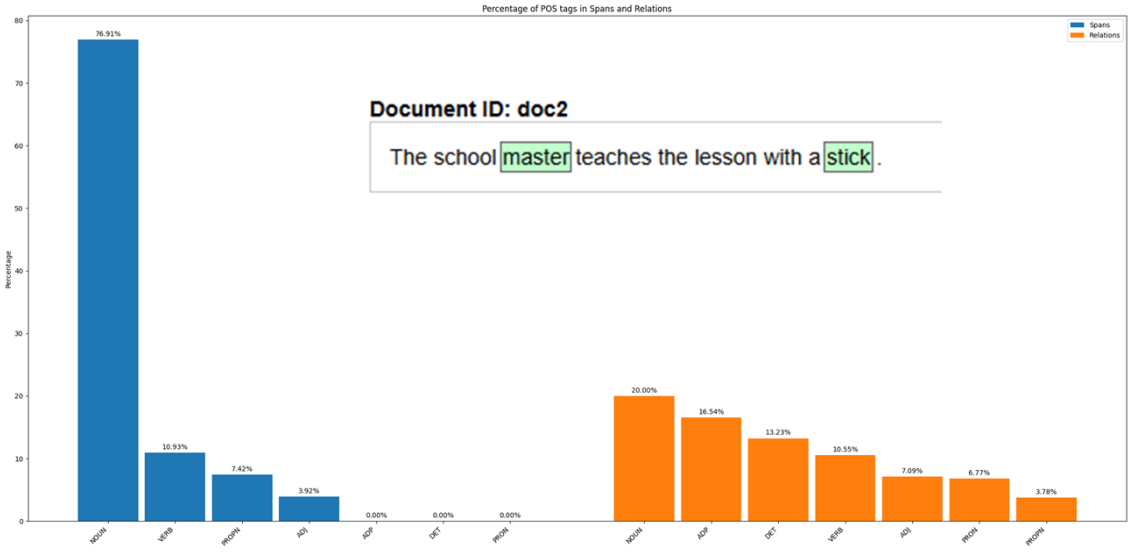


Figure 24: Semeval POS Breakdown

4.2.4 Because [51]

Somewhat similar in form to Altlex with the annotated spans a little closer to the desired format with spans representing events/states (see following POS breakdown). However this dataset still has excessive annotation noise to be usable and would need re-annotation.

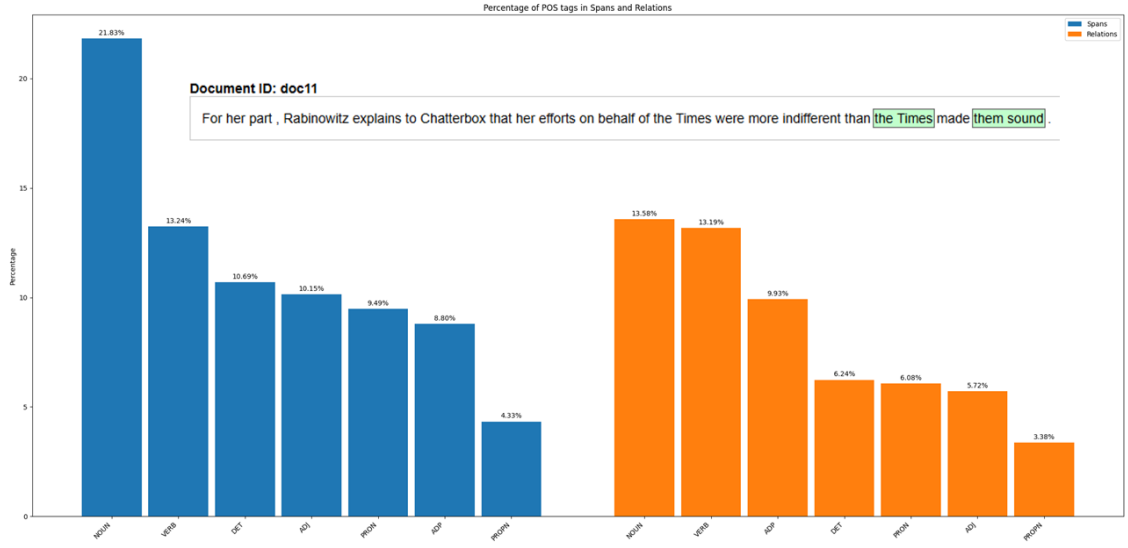


Figure 25: Because POS Breakdown

4.2.5 CTB [52]

CTB seems to have short span annotations with a mix of verbs and nouns (see following POS breakdown), but the annotations do not make a lot of sense when viewed as representing events or states.

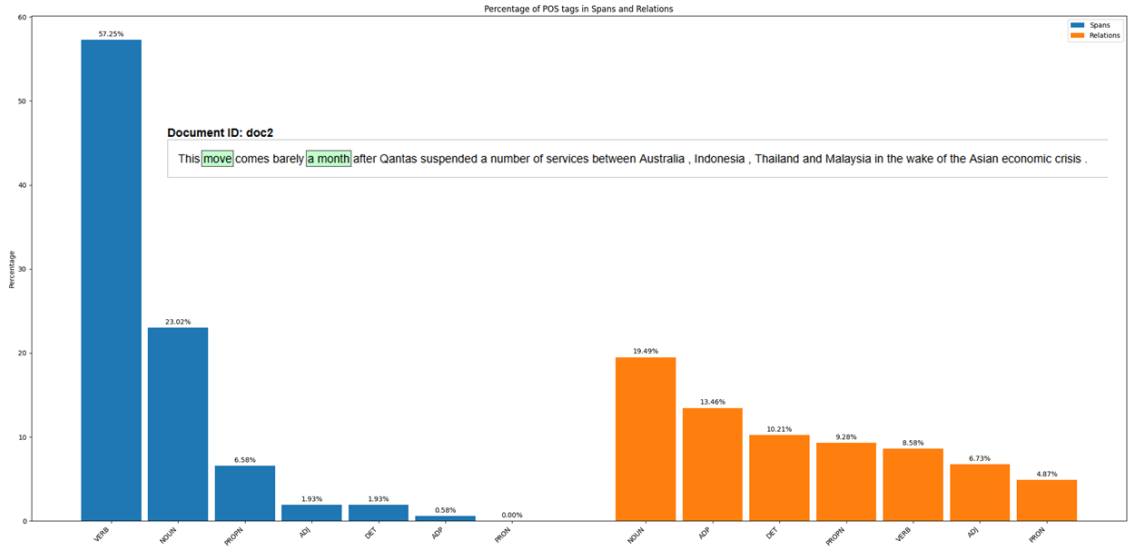


Figure 26: CTB POS Breakdown

4.3 Annotating a Custom Dataset

It was decided that in order to demonstrate the ability of the designed model to detect causal triplets, it would be best to simply annotate a custom dataset in-line with described model architecture. Chunks of observations from Altlex, Maven [56], Cassie [57], SemEval were gathered. Additionally some synthetic sequences produced by LLM prompting were added. The total observation count was approximately 900 observations. Primary issues with this dataset are the small size and the annotation noise. There is noise due to the single annotator and the short time taken to annotate. However there is also noise inherent to the ambiguous nature of annotating flat long spans representing events and states. It became clear during the annotation process that span boundaries become less defined the longer and more complex the text becomes. The maximum sequence length for this dataset was 200 words and the maximum span width was 80 words.

The annotations were done in an initial pass then used to train the model and which was then used to predict the correct annotations. These predictions were then used to review the human annotators inputs. While this was done for efficiency, it showed how the model could be useful as a first pass annotation guide.

Fig 27 shows the POS breakdown of the spans as well as the relation context tokens.

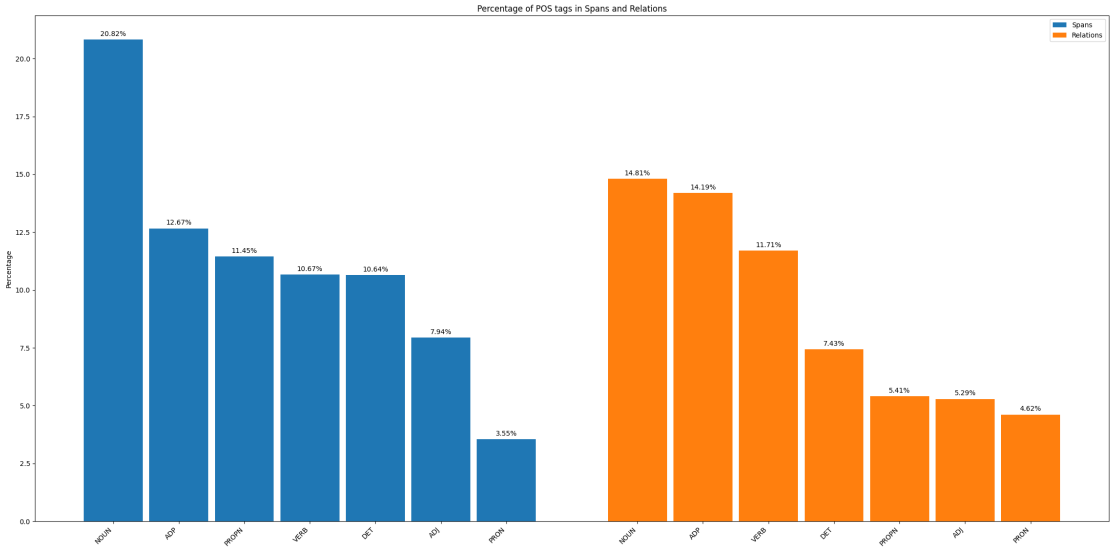


Figure 27: Custom Dataset POS Breakdown

Fig 28 gives some examples of the annotated observations from this dataset.

Document ID: doc0

An English textile industry was established in the 15th century , providing the basis for rapid English capital accumulation .

Document ID: doc1

In November 2006 , Lorin Maazel submitted Barenboim 's name as his nominee to succeed him as the New York Philharmonic 's music director .

Document ID: doc2

Many of these deaths were caused by war crimes committed by German and Japanese forces in occupied territories .

Document ID: doc3

Von Neumann acknowledged that the central concept of the modern computer was due to this paper .

Document ID: doc4

Consequently , the U.S. could find itself bombing operational missiles were the blockade to fail to force Khrushchev to remove the missiles already on the island .

Document ID: doc5

Low sea-levels meant that Britain was still attached to the continent for much of this earliest period of history , and varying temperatures over tens of thousands of years meant that it was not always inhabited at all .

Figure 28: Custom Dataset Examples

4.3.1 Annotation Guidelines

These annotation guidelines define a flat span-based approach for identifying events and states in text. This schema was chosen because it aligned well with the model architecture—specifically, a span classification framework that predicts full event spans directly. It also proved to be far more practical and intuitive than complex hierarchical trigger-argument-role schemes, especially for solo annotation. By selecting a single contiguous span that captures the event or state along with its relevant arguments, this method simplifies both the annotation process and the model's learning task.

That said, this approach is not without challenges. In sentences with convoluted struc-

ture—such as causal triggers surrounded by noisy or scattered arguments—it can be difficult to decide on clear span boundaries. These edge cases highlight some limitations of the flat span method, but in the majority of situations, it remains a practical and scalable solution that balances expressiveness with simplicity.

Flat Schema

- Each span represents a **single coherent event or state**.
- No nested structures or trigger/argument decomposition.
- Annotate the **best contiguous span** that expresses the event/state.

Span Scope: Include Relevant Arguments

- Include the event trigger and all relevant arguments (e.g., agent, theme, time, cause).
- Minor interleaved noise is acceptable if it improves span coherence.
- **Example:** *"A few months after the hotel's bombing the Government of Pakistan had reconstructed it"* is a valid single event span.

Span Boundary Preference

- Prefer the **widest natural span** that fully describes the event or state.
- Avoid fragmenting semantically integrated components (e.g., time or agent).

Annotating Embedded Events Separately Annotate embedded events as separate spans if:

- They are **independently eventive**.
- They could be the **head of a causal relation**.
- They are **referable** elsewhere in the text.
- They convey meaning **beyond acting as a modifier**.

Overlapping Spans Are Allowed

- Overlap is permitted if multiple distinct events share part of the same text.
- Flat schema means no hierarchy, but does not require disjoint spans.

What Not to Annotate

- Do **not** annotate attributional or discourse-level elements (e.g., *"according to Xinhua"*).
- Do **not** annotate standalone time expressions unless eventive.
  - **Yes:** *"the explosion on 5 May"*
  - **No:** *"in 2008"*

Causal Relations

- Annotate binary causal links between spans only if:

- One event **logically or directly causes** the other.

- Avoid speculative, metaphorical, or weakly implied causality.

Heuristic for Ambiguous Cases *"Would this text span make sense as a row in a knowledge graph?"*
If yes — annotate it.

4.4 Model Configuration Space

Rather than defining distinct model variants, a single modular pipeline was developed to support configurable architectural components. Most experiments were conducted by toggling or swapping individual modules, allowing controlled analysis of their impact.

Key configurable components included:

- **Backbone Model:** Swappable HuggingFace transformer (e.g., BERT base cased, SpanBERT base cased)
- **Subtoken Pooling Strategy:** Max-pooling (default), but framework allows alternate strategies
- **Span Generator and Filter:** Brute-force generation with binary filtering vs. token-level boundary tagging
- **Span Representation Strategy:** how the span representations were constructed from the word token embeddings, this was configurable and modular
- **Relation Representation Strategy:** how the relation representations were constructed from the word token embeddings, this was configurable and modular
- **BERT Sharing:** Shared vs. separate encoders for span and relation modules
- **BiLSTM Module:** Optional 3-layer BiLSTM over contextualized token representations
- **Graph Module:** Optional multi-head self-attention (transformer-style)
- **Projection Head:** FFN with expansion layer + ReLU, dropout vs. single linear layer
- **Top-K Spans:** the number of shortlisted spans from the span filter
- **Top-K Relations:** the number of shortlisted relations from the relation filter
- **Span/Rel Marking:** Briefly tested marked-sequence re-encoding through BERT; found too slow and ineffective

Common configurations across most variants:

- **Backbone:** BERT base cased or SpanBERT base cased
- **Subtoken Pooling:** Max-pooling
- **Backbone Sharing (Spans and Relations):** Enabled
- **Dropout:** 10%
- **Max Sequence Length:** Up to 200 word tokens
- **Max Span Width:** 80 word tokens
- **Projection Layer Type:** FFN with an intermediate expansion layer and nonlinearity, used instead of a single linear layer
- **BiLSTM:** 3 layers
- **Span Types:** Uniclass (each span assigned one of N types)

- **Relation Types:** Multilabel (each relation may take multiple of M types)
- **Span Representation Strategy:** `start + end + maxpool(inner) + width emb + [CLS]`, followed by FFN re-projection
- **Relation Representation Strategy:** `head + tail + cross-attn(head,context) + cross-attn(tail,context)`, followed by FFN re-projection
- **Span Width Embedding Size:** 100
- **Span Start/End Context Window Size:** 20% of span width
- **Redundant Span Pruning Threshold:** 80% overlap
- **Fallback for Missing Context Tokens:** Learned embedding
- **Relation Context Window (before/after):** 30 tokens
- **Top-K Spans:** 30
- **Top-K Relations:** 50
- **Graph MHA Transformer:** 3 layers, 8 heads
- **Relation Head Prediction Threshold:** 0.3

4.5 Training Configuration

The model was trained using the AdamW optimiser [58] with a learning rate of 1×10^{-5} for the pre-trained transformer encoder and 5×10^{-5} for all other parameters. Weight decay was set at 0.01.

A linear scheduler was used with a total of 20,000 steps (batches), including a warm-up phase of 2,000 steps.

Floating-point precision was maintained at full. Gradient accumulation was not employed. Gradient clipping was applied with an initial threshold of 10, linearly reduced to a minimum threshold of 1 halfway through training (at step 10,000).

The training batch size was set to 2, while inference utilized a batch size of 4. The terms 'batch' and 'step' are used interchangeably. The training set was shuffled, and the loader continuously cycled through batches until explicitly stopped. Training progress was controlled by specifying the number of steps rather than epochs.

The evaluation loop consisted of a single pass through the validation split. During training, evaluation was typically performed every 100 steps: training paused, and the model ran a full evaluation cycle over the validation split in inference mode, computing performance metrics and validation loss. These validation metrics were used to determine when to halt training. Additionally, a randomised evaluation interval was introduced to prevent periodic evaluation biases (see Section 5.3.2 on early stopping).

4.6 Loss Structure

While the model could be trained using only the span and relation classification losses, performance improved significantly when additional auxiliary losses were introduced, providing the model with greater flexibility to learn. The total loss was defined as the weighted sum of the following components:

- Token tagger or span filter loss
- Lost relation (from missing span) penalty
- Relation filter loss
- Graph structure Loss
- Span classification loss
- Relation classification loss
- Redundant Span Penalty, Hanging Relation Penalty

4.6.1 Warm-up Teacher Forcing

In all model variants, teacher forcing was implemented in the span and relation filtering stages. Specifically, logits (scores) for positive cases were artificially set to a large positive number, ensuring these cases always ranked at the top of the shortlisted K spans or relations. The remaining shortlisted cases thus served as negative samples that the model predicted as the most likely false positives. No other forms of teacher forcing were employed in the model pipeline.

It was observed that disabling this teacher forcing strategy after an initial warm-up period was beneficial for model performance, provided that an auxiliary penalty was introduced. This penalty, termed the *lost relation penalty*, penalised the model whenever a missed span prevented a positive relation from entering the final set of relation candidates. This effect can be clearly observed in Figure 35, where the teacher forcing is disabled at step 3000, triggering an immediate rise in the lost relation penalty. Subsequently, the penalty steadily decreases as the model improves its ability to identify correct spans.

4.7 Implementation and Hardware

The models were implemented in PyTorch using the HuggingFace Transformers library. Training was performed on T4 or L4 GPUs, using full-precision. Logging included evaluation every 50 steps, with best checkpoint saved based on a validation F1 related metric.

5 Results and Discussion

5.1 Quantitative Performance

Model performance is summarized in Table 3, with best configurations for each dataset in Table 4 and complete configuration details in Appendix A:. Additionally, the F1 performance is visualised in Figures 29 and 30. The following are some general trends observed from the data:

- For all model configurations, the span and relation F1 variance increases for the custom dataset, this is very likely due to the longer spans and more ambiguous boundaries. Additionally the custom dataset was not cleaned in any way.

- Relation F1 is always below span F1, this is not suprising for this class of model or data as relations are constructed from spans, thus to identify an relation you need to identify two spans.

- Conll04 and SemEval08 had far better numbers than the custom dataset, again, most likely due to the shorter spans and less associated noise.

- Spanbert-base-cased worked better on the custom dataset, while bert-base-cased worked better on the short span datasets (Conll04 and SemEval08). Again this is unsuprising given that the custom dataset consists of much longer spans. Additionally bert cased worked better than bert uncased, which is expected as no cleaning was done on the datsets.

- Using separate transformer encoders for the spans and relations paths, did not have any benefit with a shared backbone giving better results.

- Token Tagging with a BE (Begin, End) scheme was close in performance to the brute force method. The BE scheme was slightly better than the BECO (Begin, End, Combined, Other) scheme. The main advantage of token tagging was that the token tagging methods used far fewer compute resources than the brute force method.

- Attention pooling to form the span representations was inferior to the more simple method of concatenating max-pooled representations from the start window, inner window and end window along with the width embedding and the cls token from bert/spanbert. The attention pooling used excessive resources also.

- Relation representation structure appeared to prefer the concatenation of the head, tail and context representations. With the head and tail being similar to the span representation minus the width embedding and cls embedding.

- For each relation, the context tokens were either the tokens between the spans and optionally the tokens in a window before and after the head and tail spans respectively. The pooling of these tokens was either max-pooling or cross attention pooling using the head/tail representation. The cross attention pooling over the between and window context tokens performed best for the custom dataset. Max-pooling between only worked best for the short span datasets. This makes sense considering the more complex signals in the custom dataset, thus the cross attention over a wider area may have had an easier time finding it.

- Using a lower span top K (30) for the span filter seemed to work better than a larger top K. The span top K greatly affects the resource utilisation as the number of candidate relations that need a representation are the square of the span top K. The span filter stage seemed to be good at finding the best spans and allowed the use of lower top K numbers. Potentially lower numbers could have been tested, the main limit is the maximum number of spans per observation. The relation top K was tested at 200 and 50 with no noticeable difference, so it was left at 200. The relation top K has less impact on resource usage.

- Disabling of both the LSTM after the transformer encoder and the graph transformer after the relation filter. The performance appeared to favour using the LSTM and graph transformer, although the differences were not great. Both of these structures worked by further enhancing the incoming representations (token representations for the LSTM, span/relation representations for the graph transformer). They did not add that much extra resource usage, so they were just left in for most configurations.

Note: No targeted analysis was conducted to isolate which types of examples benefited from the graph transformer (e.g., complex chains or overlapping relations); evaluation was based solely on overall F1 metrics and training dynamics.

- Warm-up teacher forcing appeared lead to better results as opposed to always on teacher forcing. The warm-up period was anywhere from 1000-3000 steps. Reasons for this could be that running the data through the model in training without labels allowed the model to learn under similar conditions to inference.

- Three penalties were used:
    1. lost relations (caused by the span filter missing spans)
    2. redundant spans (caused by the span classification head classifying very similar spans
    3. hanging relations (caused by the relation classification head classifying relations between one or both spans that did not make it through the span classification head

  Each one of these penalties was tied to the logit of the offending source span/relation so it would nudge the model. The penalties were enabled when teacher forcing was disabled. The penalties appeared to improve the results somewhat, however, it should be stated that the model already has quite a few avenues to learn from and it works fine without these penalties, it may just take longer to learn.

- Other aspects of the model not mentioned in these experimental results due to time constraints and uninteresting performance were:
    - Marking spans and relations and re-running through bert showed no noticeable benefit to performance. Additionally it was extremely resource intensive.
    - Cosine similarity loss as opposed to regular strict losses was not a viable training option, giving no usable signal to the model to learn from.

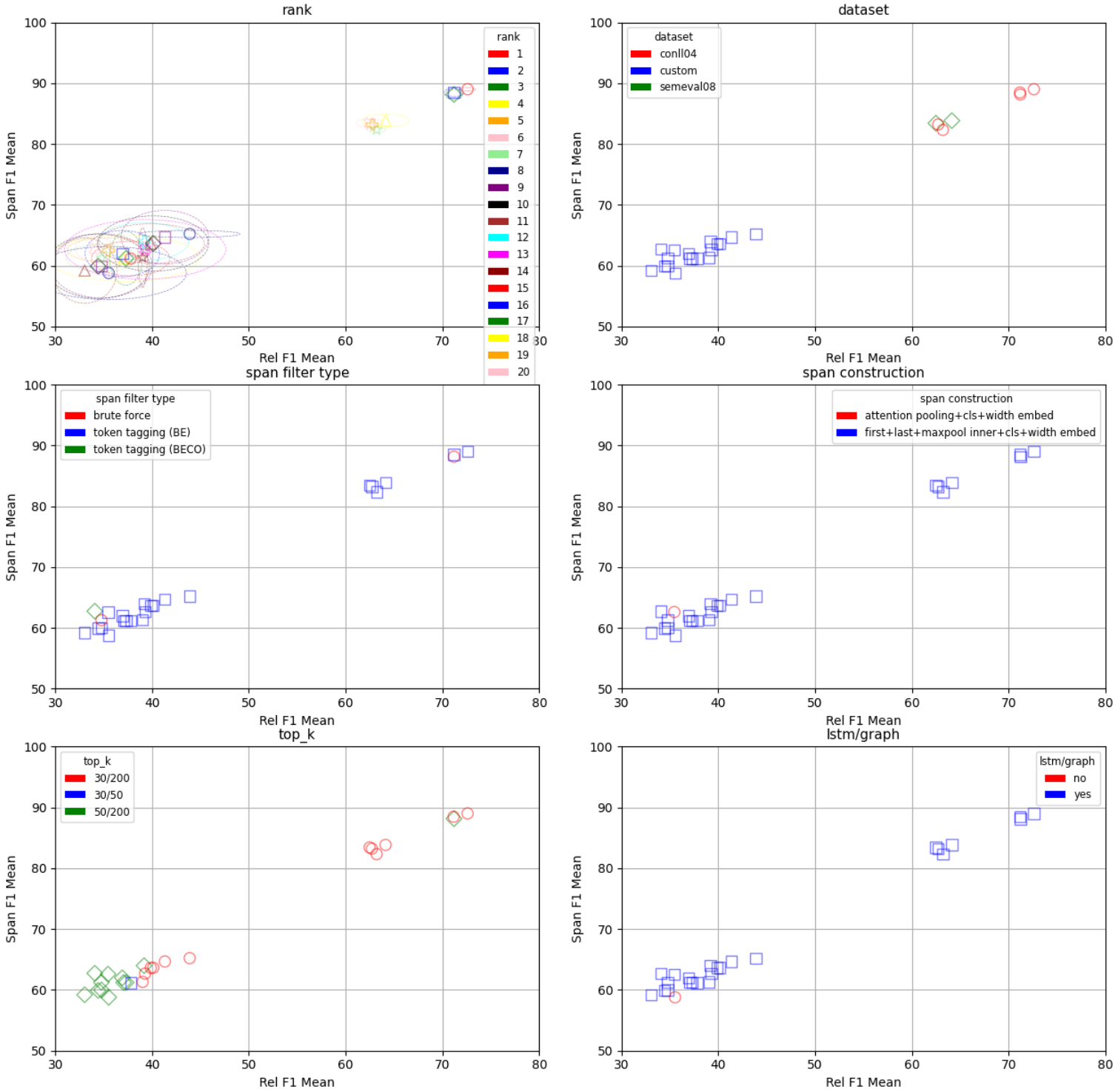


Figure 29: Experiment Results Part 1

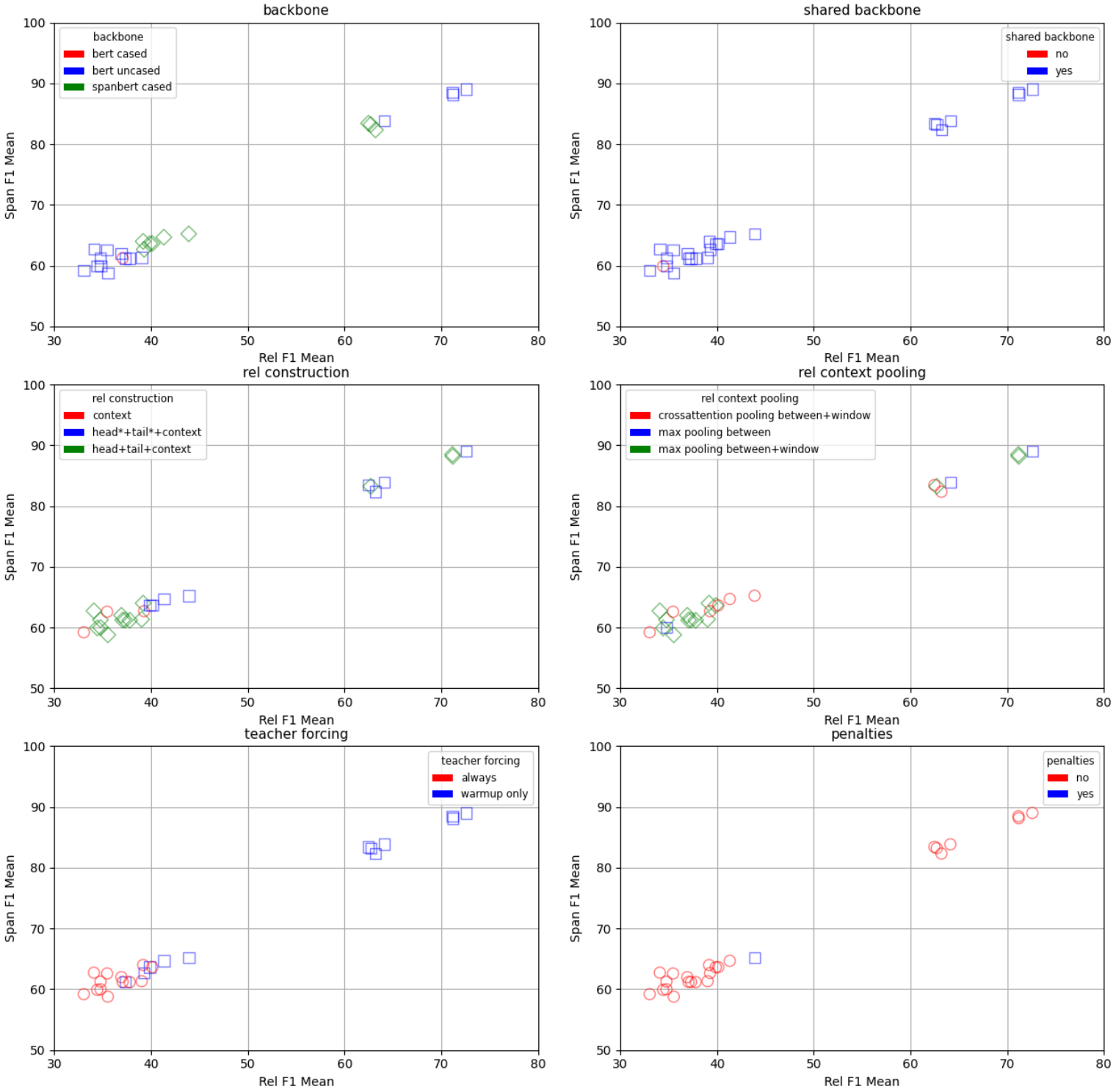


Figure 30: Experiment Results Part 2

Model	Span P	Span R	Span F1	Rel P	Rel R	Rel F1
conll04_1	89.11 ± 0.96	87.24 ± 1.26	88.15 ± 0.83	72.64 ± 2.61	69.87 ± 1.61	71.19 ± 0.92
conll04_2	$\mathbf{89.36 \pm 0.50}$	$\mathbf{88.69 \pm 1.01}$	$\mathbf{89.02 \pm 0.26}$	$\mathbf{73.10 \pm 0.57}$	$\mathbf{72.11 \pm 1.35}$	$\mathbf{72.60 \pm 0.82}$
conll04_3	89.23 ± 0.89	87.76 ± 0.83	88.48 ± 0.68	70.52 ± 0.73	71.82 ± 1.66	71.16 ± 1.03
conll04_4	82.79 ± 2.09	81.93 ± 1.18	82.34 ± 0.46	65.48 ± 3.38	61.23 ± 2.54	63.20 ± 0.94
conll04_5	84.71 ± 2.04	81.82 ± 0.81	83.23 ± 0.63	62.77 ± 1.09	62.68 ± 1.05	62.71 ± 0.21
semeval08_1	$\mathbf{83.61 \pm 2.87}$	$\mathbf{84.15 \pm 1.08}$	$\mathbf{83.85 \pm 1.08}$	$\mathbf{65.43 \pm 1.76}$	$\mathbf{62.90 \pm 3.51}$	$\mathbf{64.12 \pm 2.44}$
semeval08_2	83.43 ± 2.11	83.48 ± 0.81	83.43 ± 0.69	62.81 ± 1.68	62.20 ± 2.44	62.47 ± 1.37
custom_1	60.66 ± 3.42	61.78 ± 4.22	61.21 ± 3.80	41.23 ± 11.41	33.89 ± 5.79	37.06 ± 7.78
custom_2	58.91 ± 5.96	64.10 ± 2.18	61.26 ± 2.24	34.30 ± 0.21	35.61 ± 7.33	34.77 ± 3.66
custom_3	62.00 ± 2.39	63.20 ± 1.60	62.59 ± 2.01	34.33 ± 4.44	36.70 ± 4.24	35.45 ± 4.22
custom_4	58.32 ± 5.79	60.13 ± 5.08	59.20 ± 5.39	37.55 ± 5.55	29.55 ± 2.14	33.03 ± 3.33
custom_5	62.10 ± 3.93	61.94 ± 4.11	61.99 ± 3.79	38.31 ± 7.73	36.60 ± 4.07	36.92 ± 3.85
custom_6	57.52 ± 5.44	60.38 ± 5.48	58.78 ± 4.36	35.94 ± 8.58	35.17 ± 7.59	35.52 ± 8.01
custom_7	57.74 ± 4.20	62.27 ± 6.53	59.91 ± 5.29	35.83 ± 3.70	33.24 ± 6.63	34.43 ± 5.32
custom_8	60.56 ± 5.28	61.92 ± 3.85	61.21 ± 4.43	37.39 ± 3.44	37.35 ± 0.88	37.31 ± 1.69
custom_9	60.24 ± 5.84	59.74 ± 4.83	59.98 ± 5.29	35.54 ± 6.60	34.10 ± 3.94	34.77 ± 5.24
custom_10	61.51 ± 6.35	61.29 ± 4.43	61.33 ± 4.89	41.38 ± 0.35	36.92 ± 0.97	39.02 ± 0.48
custom_11	61.79 ± 6.29	60.85 ± 1.13	61.18 ± 2.85	37.86 ± 7.67	38.29 ± 1.74	37.79 ± 4.04
custom_12	61.77 ± 3.38	63.80 ± 1.85	62.73 ± 2.01	34.49 ± 2.25	33.74 ± 2.60	34.08 ± 1.98
custom_13	$\mathbf{67.25 \pm 1.70}$	$\mathbf{63.34 \pm 0.42}$	$\mathbf{65.22 \pm 0.59}$	$\mathbf{45.87 \pm 7.99}$	$\mathbf{42.19 \pm 2.79}$	$\mathbf{43.88 \pm 5.18}$
custom_14	61.27 ± 3.69	64.22 ± 6.42	62.68 ± 4.90	40.78 ± 9.73	38.23 ± 8.21	39.27 ± 8.32
custom_15	64.87 ± 2.90	62.52 ± 6.06	63.64 ± 4.55	45.76 ± 7.00	35.84 ± 5.27	40.12 ± 5.72
custom_16	63.77 ± 6.89	65.91 ± 3.45	64.68 ± 4.41	44.15 ± 6.06	39.02 ± 4.09	41.32 ± 4.53
custom_17	62.72 ± 2.75	64.58 ± 3.87	63.62 ± 3.27	42.55 ± 9.50	37.94 ± 3.00	39.87 ± 5.54
custom_18	64.28 ± 3.52	63.80 ± 3.72	63.99 ± 3.10	41.05 ± 6.97	37.72 ± 3.28	39.17 ± 4.60

NOTE: performance metrics shown are for the validation set, not the test set. This was a result of resource, time and financial constraints. As will be shown later, the correlation between validation set and test set save scores (see early stopping explanation 5.3.2) very high so it was not helpful to redo the experiments again. However, note that there is some variation so the test F1 scores would typically be 5-10% points lower than the validation set numbers.

Table 3: Micro-averaged strict precision (P), recall (R), and F1 scores for span and relation extraction.

Model	Features
conll04_2	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: max pooling between **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
semeval08_1	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: max pooling between **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
custom_13	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: crossattention pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: yes

Table 4: Best Config Details for Experiments

5.2 Ablation Studies

For investigation into some of the key features of the model, four aspects were selected for ablation studies, namely:

- Token Tagging vs Brute Force Span Search
- Relation Context Pooling Method
- Use a Graph Transformer
- Use of Penalties

The Ablation results were collected from the test set metrics on the custom long span dataset using 3 random seeds and 2 runs per seed. The seed refers to the seed used to make the train-val-test splits. The random seed used to initialise the model was always 42. The model initialisation random seed was not varied as there were so many other sources of variation in the training process, it made no difference. Repeated runs with identical seeds still resulted in different outcomes due to the unstable nature of the training process. This variance complicated both early stopping and model comparison

The baseline had the following configuration:

- backbone = spanbert base cased
- BiLSTM with 3 layers
- max span length = 200, max span width = 80 (50 for the brute force comparison)
- span representation method = start window maxpool, inner maxpool, end window maxpool, width embedding, cls embedding
- span filtering = token tagging with BE scheme
- relation representation construction = head*, tail*, context. Where * denotes span representations without width and cls embeddings.
- relation context pooling = cross attention with the head and tail spans
- top K spans = 30, top K relations = 200
- graph transformer with 8 heads and 3 layers
- using lost relation, redundant span and hanging relation penalties

One overall observation is consistent with the other experimental results is that the relation metrics have a much higher variance than the span metrics. This is not suprising due to the pipeline architecture and the relations being dependent on spans. Lastly, these ablation results highlight the sensitivity of relation extraction performance under noisy settings, and show that several architectural and training decisions, though not significant in isolation, can compound into more pronounced effects.

Model	Span P	Span R	Span F1	Rel P	Rel R	Rel F1
baseline (sw 50)	59.06 ± 1.53	59.88 ± 1.48	59.46 ± 1.25	37.84 ± 5.06	36.25 ± 7.47	36.64 ± 5.36
brute force spans	$\mathbf{59.58 \pm 1.58}$	$\mathbf{61.31 \pm 2.11}$	$\mathbf{60.41 \pm 1.39}$	$\mathbf{40.29 \pm 5.85}$	$\mathbf{37.7 \pm 5.06}$	$\mathbf{38.89 \pm 5.16}$
baseline	$\mathbf{60.25 \pm 1.97}$	$\mathbf{59.74 \pm 1.74}$	$\mathbf{59.98 \pm 1.58}$	$\mathbf{38.76 \pm 7.58}$	$\mathbf{37.79 \pm 7.01}$	$\mathbf{38.15 \pm 7.03}$
no graph	58.13 ± 1.28	59.89 ± 1.59	58.98 ± 1.09	35.65 ± 6.91	36.07 ± 6.73	35.68 ± 6.18
graph 12h-6l	56.69 ± 2.29	58.77 ± 2.57	57.64 ± 0.88	35.89 ± 8.22	36.93 ± 5.06	36.27 ± 6.56
graph 16h-8l	58.13 ± 1.41	58.34 ± 2.48	58.22 ± 1.69	32.98 ± 6.32	35.21 ± 7.67	34.01 ± 6.86
graph 32h-16l	58.06 ± 2.56	58.89 ± 1.99	58.45 ± 2.0	35.36 ± 5.7	38.56 ± 4.87	36.87 ± 5.25
baseline	$\mathbf{60.25 \pm 1.97}$	$\mathbf{59.74 \pm 1.74}$	$\mathbf{59.98 \pm 1.58}$	$\mathbf{38.76 \pm 7.58}$	$\mathbf{37.79 \pm 7.01}$	$\mathbf{38.15 \pm 7.03}$
no consist penalties	57.12 ± 2.23	59.28 ± 1.54	58.16 ± 1.61	35.88 ± 8.06	38.14 ± 8.27	36.96 ± 8.15
no lr penalties	56.99 ± 3.67	59.25 ± 3.33	58.02 ± 2.63	37.2 ± 6.68	36.22 ± 5.68	36.54 ± 5.68
no penalties	57.08 ± 3.39	58.28 ± 1.68	57.64 ± 2.29	31.95 ± 3.27	34.88 ± 6.82	33.25 ± 4.6
baseline	$\mathbf{60.25 \pm 1.97}$	$\mathbf{59.74 \pm 1.74}$	$\mathbf{59.98 \pm 1.58}$	$\mathbf{38.76 \pm 7.58}$	$\mathbf{37.79 \pm 7.01}$	$\mathbf{38.15 \pm 7.03}$
rel cxt crossattn 16h	60.39 ± 3.22	61.17 ± 2.26	60.76 ± 2.59	35.67 ± 8.09	37.12 ± 7.59	36.17 ± 7.35
rel cxt maxpool	57.56 ± 2.13	60.22 ± 2.89	58.84 ± 2.21	35.96 ± 6.12	37.41 ± 3.99	36.44 ± 4.34

Table 5: Micro-averaged strict precision (P), recall (R), and F1 scores for span and relation extraction across ablation groups for the test set.

5.2.1 Token Tagging vs Brute force Span Search

Two competing span filtering methods where evaluated, token tagging vs brute force span search. The advantage of token tagging is that is could be scaled to much larger sequence lengths and span widths as was far more resource efficient than the brute force method. In this comparison the tagging scheme used in the baseline was BE (Begin-End). For the brute force setup, span representations were made for all possible spans given the batch max sequence length and max span width. An A100 processor was required given the GPU requirements of the span representations tensor and even then it could only support a max span width of 50. The span representations tensor is passed through a binary filter head and the top K spans are chosen based on the logit score. The primary advantage of the brute force method over the token tagging is that it is less likely to miss spans.

Figure 31 shows the comparison. The brute force method achieved slightly better test set metrics than the baseline, although the range of values was broadly similar across both. In short, brute force is preferable when feasible (e.g., NER-RE tasks), but impractical for longer spans typical of EE-RE tasks.

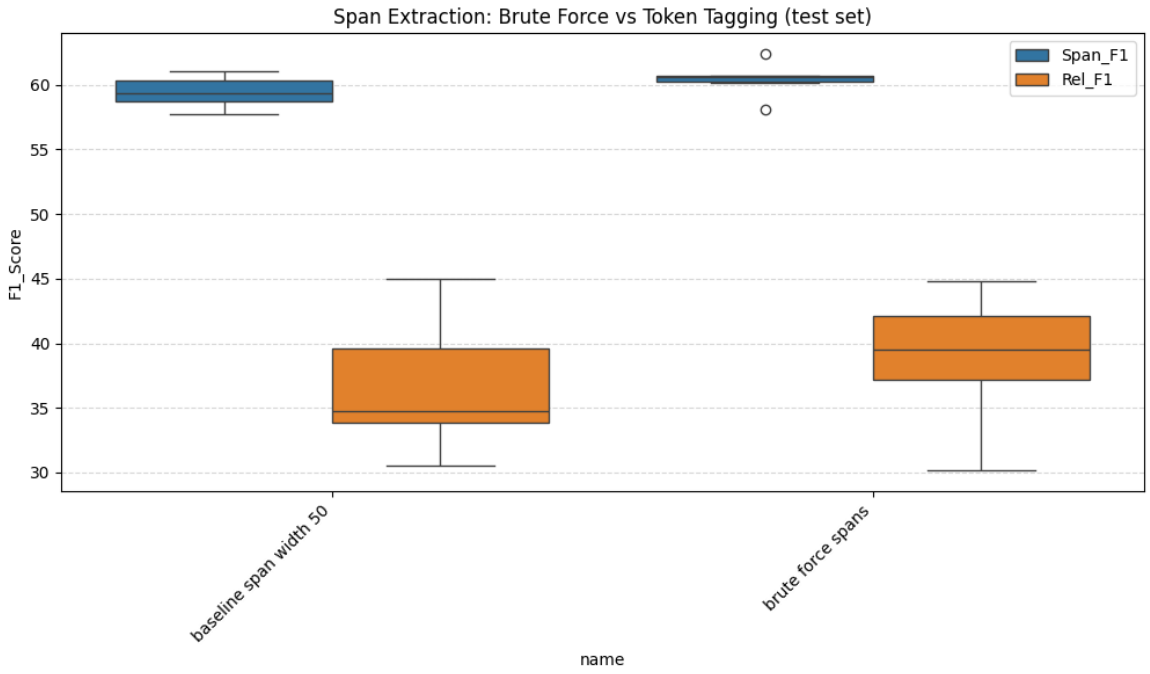


Figure 31: Token Tagging vs Brute Force Span Search

5.2.2 Relation Context Pooling Method

This compared the method for pooling the variable number of context token embeddings into a context representation used to construct the relation representation. The baseline uses the concatenation of the cross-attention with the head and tail span representations. Each cross-attention block uses the head or tail span representation as query and the context token embeddings as key, value. The first variant tested was using 16 heads for this MHA block as opposed to the baseline of 8 heads. The second variant tested used much simpler max-pooling over the context tokens.

Figure 32 shows minimal performance differences between methods. The baseline appears marginally stronger overall. Max-pooling tended to underperform in relation F1, though occasional poor runs were also observed for the baseline. This is consistent with earlier observations relating to noise in the annotation process, span boundaries, and causality signals.

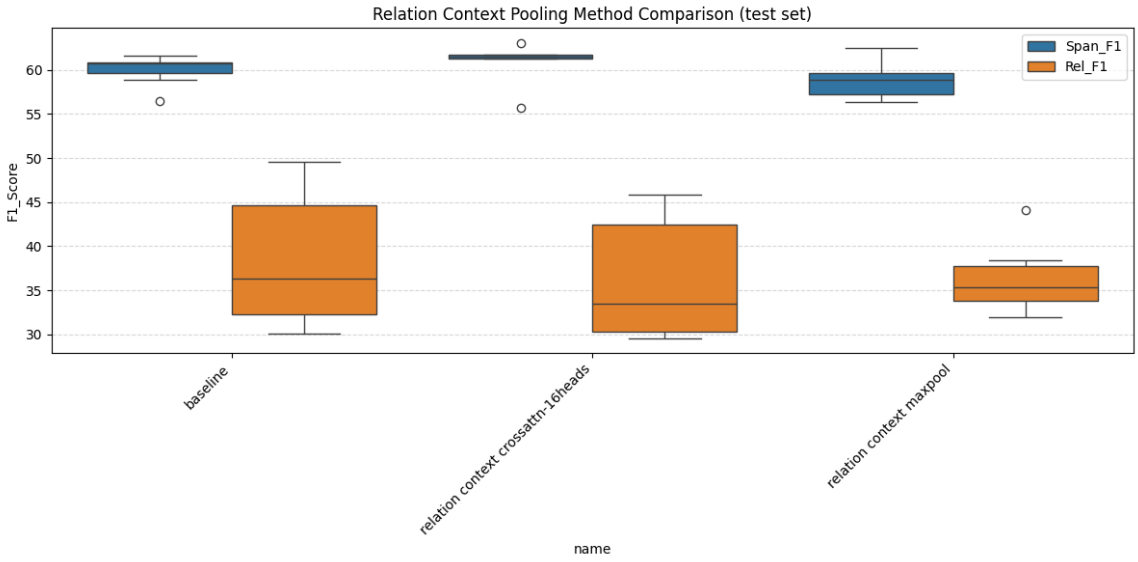


Figure 32: Relation Context Pooling Method

5.2.3 Graph Transformer

The graph transformer acts as an optional enrichment module for span (node) and relation (edge) representations before final classification. Validation results indicated a possible small benefit, and resource usage was low enough to justify its inclusion.

The ablation compared the baseline configuration (8 heads, 3 layers) against both a no-graph setup and larger graph variants. As shown in Figure 33, using a graph transformer had a marginal benefit over not using one. However, increasing its size beyond the baseline did not yield further gains and sometimes degraded performance.

In short, a lightweight graph transformer may offer small improvements at low cost, but larger configurations do not appear beneficial.

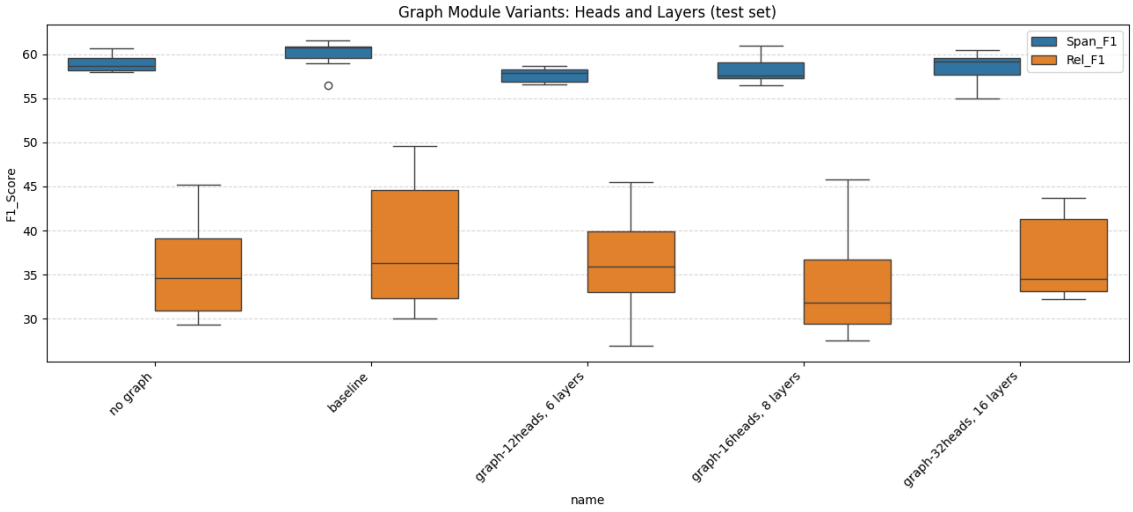


Figure 33: Graph Transformer

5.2.4 Use of Penalties

Learnable penalties were compared against the baseline which used both lost relation penalties as well as consistency penalties (redundant span, hanging relation). The results would appear to favour using both types of penalties which effectively discourage the model from making common span-relation consistency errors and span filtering errors.

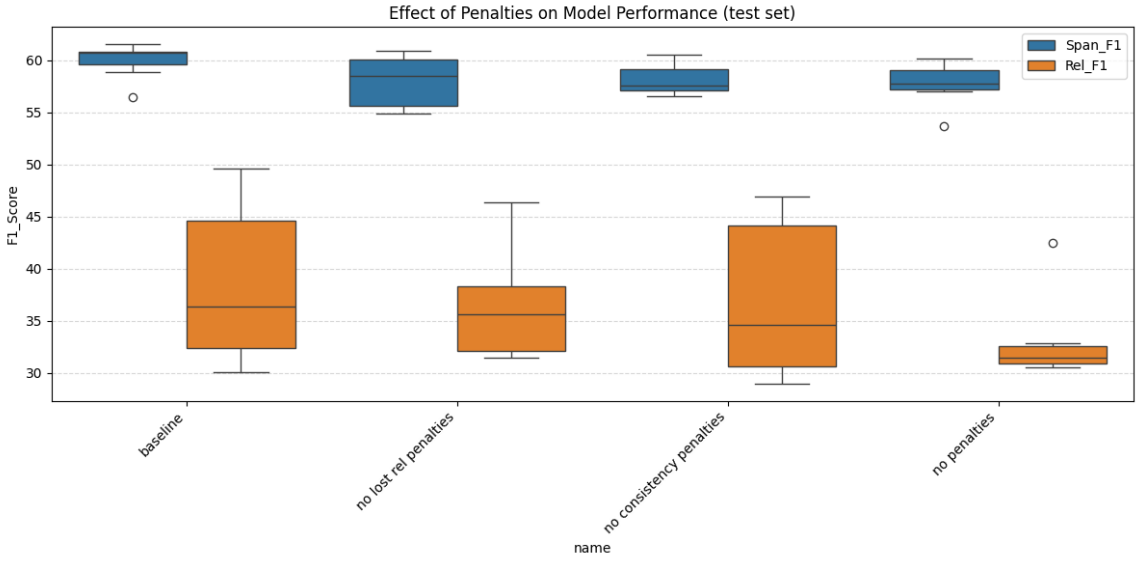


Figure 34: Penalties

5.3 Training Dynamics

5.3.1 Train Loss

As is shown in the model diagrams, 20,21 multiple losses are generated along the model pipeline in order to give the opportunity for different aspects to send training signal to the gradients. The following figure (35) shows the various training loss components changing during a training run.

NOTE: the discontinuity at step 3000 is due to the disabling of teacher forcing and the enabling of penalties

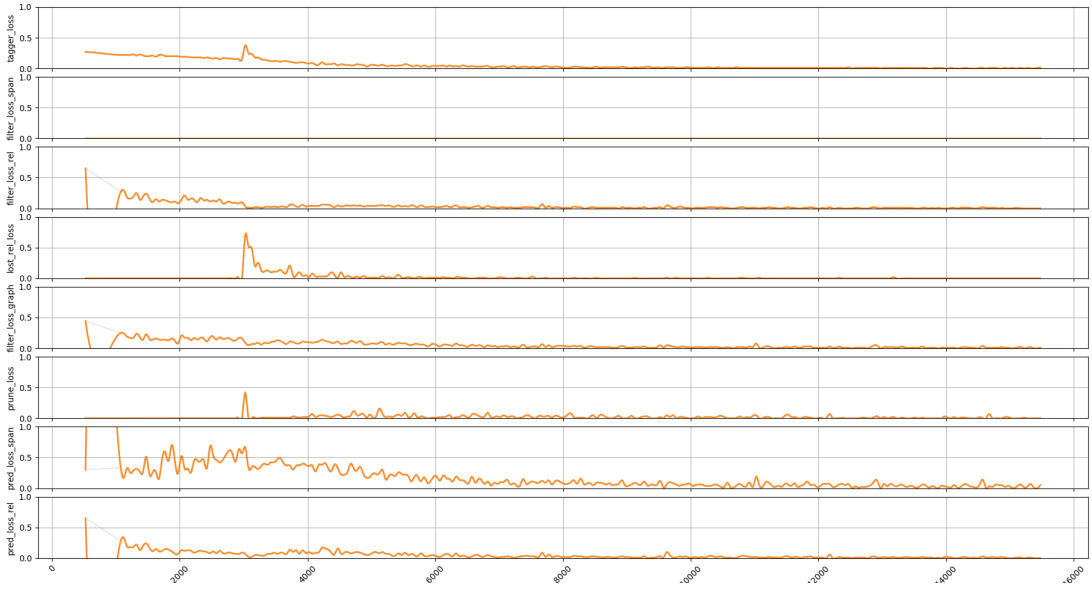


Figure 35: Train Loss Breakdown vs Train Step

5.3.2 Early Stopping

This particular model being a more complex joint loss model with multiple interacting stages appeared to have some decoupling between validation set loss and validation set metrics (precision, recall, f1). As shown in Fig 36 and 37 the train loss decreases throughout the training, while the validation set loss hits a minima while still warming up and before the model effectively starts predicting relations. As a result the training was ended not by looking for a minima in the validation set loss, but instead, looking for a maxima in a performance metric. This metric (save score) is simply the addition of the span and relation F1 with each one moderated by a balance factor as show below:

$$\textbf{Save Score} = \textbf{SpanF1} \cdot \left(\frac{\min(\textbf{SpanP}, \textbf{SpanR})}{\max(\textbf{SpanP}, \textbf{SpanR})} \right)^x + \textbf{RelF1} \cdot \left(\frac{\min(\textbf{RelP}, \textbf{RelR})}{\max(\textbf{RelP}, \textbf{RelR})} \right)^x$$

where x penalizes P/R imbalance. ($x = 2$ was used)

Additionally, the period at which the trainer checks the validation set save score was randomised around a given period (typically 50 steps). This helped to find a good stopping point.

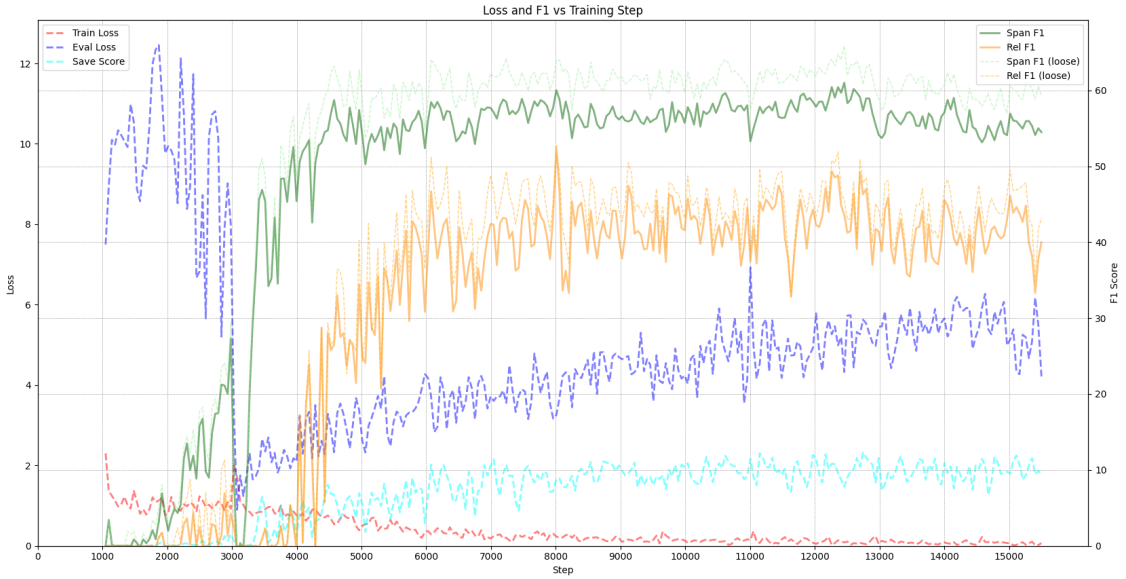


Figure 36: Training Results

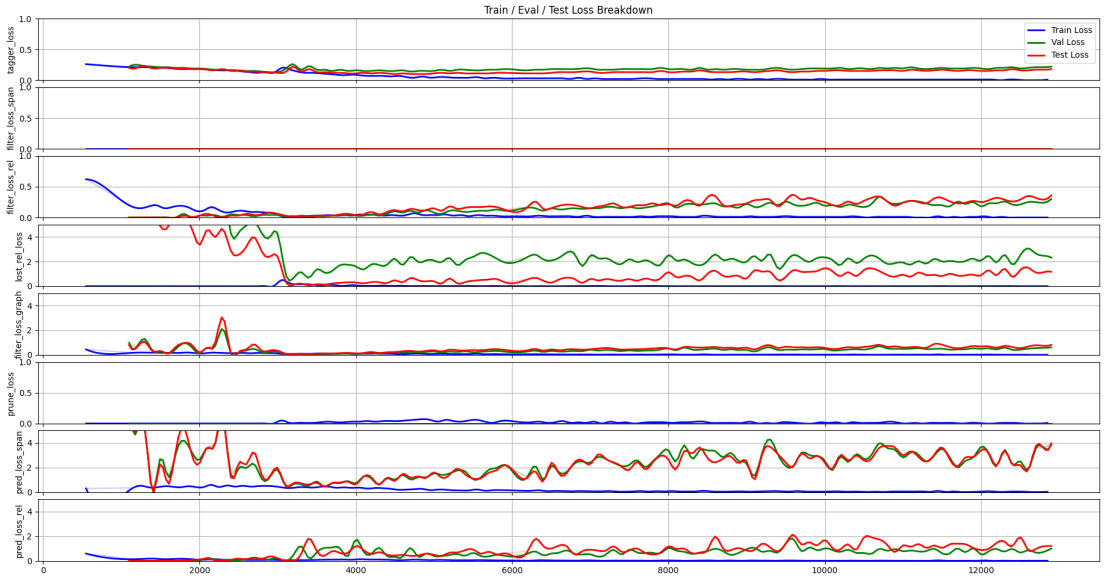


Figure 37: Train/Val/Test Loss Breakdown vs Train Step

There is concern about using validation set metric peaks to choose the stopping point of the training, this is not standard practice and there is no guarantee that a spike in validation metrics will be the same for test set metrics and broader model generalisation. But this was

the only real way of finding a stopping point. The main issue was that the model would start to identify spans and the span metrics would plateau, the relation metrics however, would only start appearing after the span F1 breached 40-50%, then the relation F1 would rise and plateau more slowly with more variance. Thus the ideal stopping point would be soon after the relation F1 plateaus. Perhaps with a cleaner, larger dataset, some of these patterns may have been different and a modified early stopping algorithm developed.

To quantify potential issues with divergence of validation and test set performance, they were analysed. The scatter plot in Fig 38 and line plot 39 confirm this relationship, showing a strong positive correlation (Pearson r = 0.974) between the validation and test save scores, despite modest variance at higher performance levels.

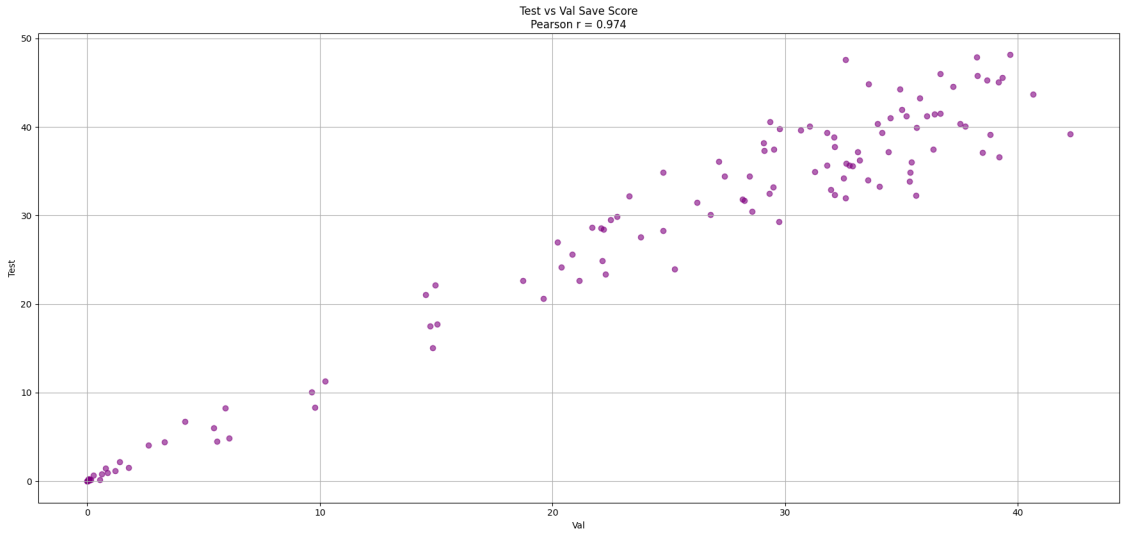


Figure 38: Test Set vs Validation Set Save Score Correlation

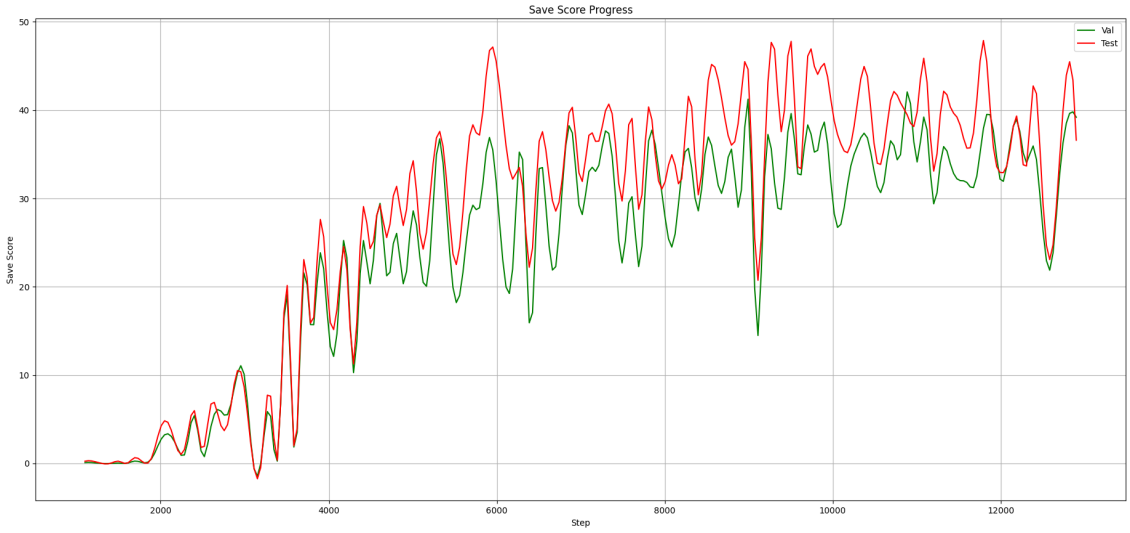


Figure 39: Test Set vs Validation Set Save Score Line Plot

5.4 Qualitative Evaluation and Error Analysis

Some human and LLM aided reviews of the model predictions vs labels were undertaken with some general outcomes indicating that the model's performance was not as poor as the F1 metrics for several reasons. The most common being poor quality annotations that did not actually abide by the annotation guidelines or were just plain wrong. Another reason is that F1 metrics can be very strict in the sense of span boundaries, while human and LLM review can be more flexible. The following is an output from one such review. Please note that due to the variability of the model and stopping point, every model run will give slightly different results.

What stands out is that in this particular model and review, for 67% of cases, the predictions were preferable or as good as the labels. The other key point is that almost 30% of cases had annotation problems.

Category	Percentage
Winner: predictions	28%
Winner: labels	32%
Winner: both	39%
Winner: neither	4%
Annotation issues flagged	28%

Table 6: Summary of evaluation outcomes.

Model Strengths:

- **Improved Span Boundary Adherence:** The model often avoids overly broad or speculative spans present in labels. It demonstrates better granularity and argument inclusion.
- **Better Causal Link Detection:** Several times, the model detects plausible causal links that labels miss due to noisy or incomplete annotations.
- **Avoids Common Label Errors:** For example, the model avoids labeling hypothetical/non-eventive spans or attributive phrases, A common mistake in labels.
- **Embedded Events:** It successfully segments embedded or compound events more accurately than the labels, in some cases.

Model Weaknesses:

- **Missed Relations:** The model occasionally fails to predict causal relations even when spans are correct. This was the most frequent reason for preferring labels.
- **Over-fragmentation or Incomplete Spans:** Some spans are too short or fragmented, excluding critical arguments or temporal markers.
- **Misinterpretation of Causality Structure:** In a few instances, causal chains are mislinked (e.g., linking cause directly to final result, skipping intermediate steps).

In summary, despite imperfections, the model exhibited useful generalization and meaningful output on ambiguous or complex examples, occasionally surpassing the reference annotations.

5.5 Example Cases

Here some of the test set cases are shown with a quick explanation of what is happening. Four examples have been chosen that demonstrate failures of both the model and the human annotator.

In Fig 40, is an example of a case with a quote, which has confused the model and potentially the annotator. The quote could have been left out entirely from the annotation with just 2 two events A) the church not granting his favour and B) Henry citing the passage in the Book of Leviticus. The model has gone for “They shall be childless”, which is not incorrect, but why is it causally related to the church event? In regards to the quote, should it be left in, should its events and states be separate spans. This shows some of the ways language can create convoluted situations that make both annotation and training simplistic models difficult.

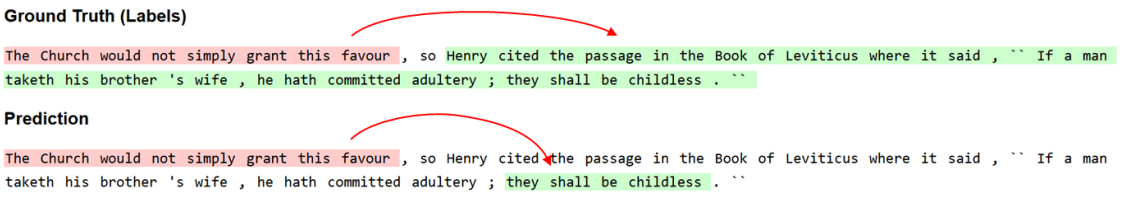


Figure 40: Example 1

Fig 41, is a generally straight forward case. Possibly the models two spans are cleaner than the annotator who has included an additional span which is really just a time attribute of the crash event, so potentially it should have been included with the crash event span, but then why did the model not include it. However, both the annotator and model got the simple causal relation.

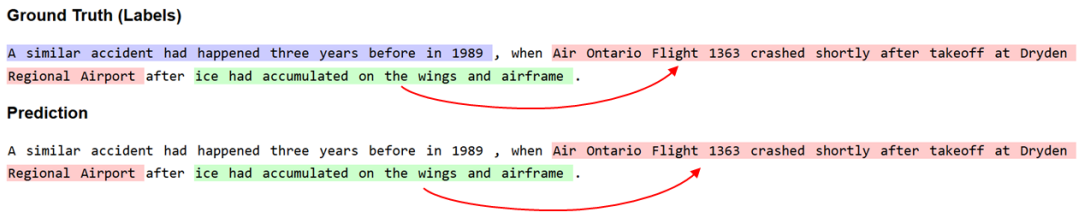


Figure 41: Example 2

Fig 42, is a longer text, which comes from a news report which includes quotes. There are many events and states that could be annotated in this text thus making it a good example of the conundrum of event and state annotation. The annotator has gone for a minimalist annotation here picking the two most obvious key causal event pairs and just ignoring anything else. The model has gone and annotated more spans, some make sense, some questionable. It has additionally missed the cyber-attack event. It did pick one good causal relation. However, neither the predictions, nor the human annotations are comprehensive in this example.

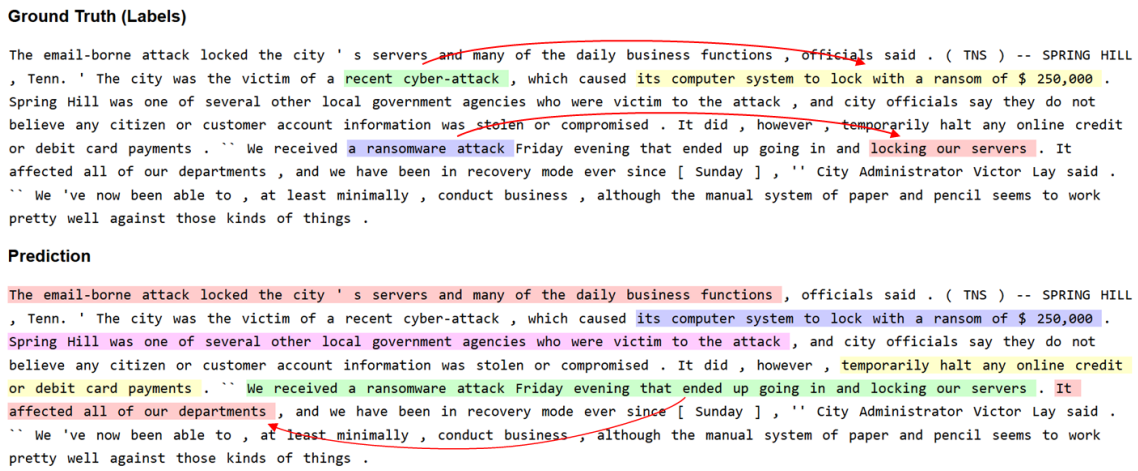


Figure 42: Example 3

Fig 43, is a moderate length sequence with much causality. The human has gone for all the contributing factors as causing “the disaster” as well as linking the worker related states to the last worker actions span. The model has come up with several overlapping spans and linked most of them back to “the disaster” as well as the relation between the worker training and the worker actions. Overall, the model had issues with overlapping spans and some questionable spans and missed relations.

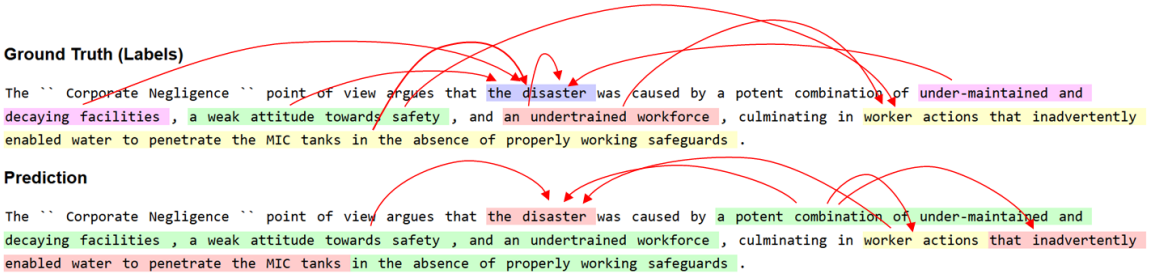


Figure 43: Example 4

6 Conclusion and Future Work

6.1 Key Findings

- Flat span based discriminative models are indeed capable of extracting useful causal structure. However, performance declines on longer spans, implicit causal signal, complex causal chains and ambiguous and convoluted language structures. To stay within the safe working zone for this class of discriminative models would give significantly better results but would require limiting the input text to shorter spans with explicit causal indicators and keeping convoluted and complex linguistic structure and semantics to a minimum as these just confuse the model and are arguably beyond the it's ability.
- The quality of training labels is a critical bottleneck — rushed or noisy spans yield poor supervision. A good metric as to how appropriate the annotation scheme is (and model architecture), would be to review the annotators feelings on the task. If the annotation guidelines are confusing, the annotators will be confused and frustrated indicating that the whole annotation strategy may be not the best way of breaking down the causal structures in the text.
- Qualitative outputs often exceeded label quality, indicating the model learned valid causal abstractions beyond strict token boundaries. Thus softer metrics or review is beneficial, this job can be done quite well with the help of LLMs.

6.2 Limitations

- The small dataset with inconsistent annotations had a performance hit for the model.
- Training and Evaluation is limited to strict span matching, missing nuanced improvements.
- This class of model appears to be limited in its ability to deal with complex language, which is expected. This is a key limitation though and is important to understand when choosing this kind of model for training.

6.3 Future Directions

- Improve annotation quality and scale via collaborative multi-pass review, and explore weakly supervised annotation techniques, which show strong potential [59].
- Investigate methods for pre-processing raw natural language into simplified and normalised, causally explicit statements more suitable for span-based models. Large generative models may be useful here, given their strengths in summarisation, rephrasing and translation.
- Evaluate generative end-to-end models (e.g., [9, 10, 28]) for the whole task.

Appendix A: Experiment Configuration Details

Model	Features
conll04_1	**backbone**: bert uncased **shared backbone**: yes **span filter type**: brute force **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
conll04_2	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: max pooling between **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
conll04_3	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no

Model	Features
conll04_4	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: crossattention pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
conll04_5	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
semeval08_1	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: max pooling between **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no

Model	Features
semeval08_2	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: crossattention pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
custom_1	**backbone**: bert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_2	**backbone**: bert uncased **shared backbone**: yes **span filter type**: brute force **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no

Model	Features
custom_3	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: attention pooling+cls+width embed **rel construction**: context **rel context pooling**: crossattention pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_4	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: context **rel context pooling**: crossattention pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_5	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no

Model	Features
custom_6	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: no **teacher forcing**: always **penalties**: no
custom_7	**backbone**: bert uncased **shared backbone**: no **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_8	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no

Model	Features
custom_9	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_10	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_11	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 30/50 **lstm/graph**: yes **teacher forcing**: always **penalties**: no

Model	Features
custom_12	**backbone**: bert uncased **shared backbone**: yes **span filter type**: token tagging (BECO) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_13	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: crossattention pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: yes
custom_14	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: context **rel context pooling**: crossattention pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no

Model	Features
custom_15	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: crossattention pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no
custom_16	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: crossattention pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no
custom_17	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head*+tail*+context **rel context pooling**: max pooling between+window **top_k**: 30/200 **lstm/graph**: yes **teacher forcing**: warmup only **penalties**: no

Model	Features
custom_18	**backbone**: spanbert cased **shared backbone**: yes **span filter type**: token tagging (BE) **span construction**: first+last+maxpool inner+cls+width embed **rel construction**: head+tail+context **rel context pooling**: max pooling between+window **top_k**: 50/200 **lstm/graph**: yes **teacher forcing**: always **penalties**: no

Table 7: Config Details for Experiments

References

[1] M. S. and, "Toward a natural language-based causal model acquisition system," *Applied Artificial Intelligence*, vol. 3, no. 2-3, pp. 191–212, 1989.

[2] R. M. Kaplan and G. Berry-Rogghe, "Knowledge-based acquisition of causal relationships in text," *Knowledge Acquisition*, vol. 3, no. 3, pp. 317–337, 1991.

[3] R. Girju and D. Moldovan, "Text mining for causal relations," in *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2002)*, pp. 360–364, AAAI Press, 2002.

[4] Q. Do, Y. S. Chan, and D. Roth, "Minimally supervised event causality identification," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, (Edinburgh, Scotland, UK), pp. 294–303, Association for Computational Linguistics, 2011.

[5] N. Asghar, "Automatic extraction of causal relations from natural language texts: A comprehensive survey," *arXiv preprint arXiv:1605.07895*, 2016.

[6] J. Frattini, M. Junker, M. Unterkalmsteiner, and D. Mendez, "Automatic extraction of cause-effect-relations from requirements artifacts," in *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 1–12, IEEE/ACM, 2020.

[7] A. Vaswani *et al.*, "Attention is all you need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, (Long Beach, CA, USA), 2017.

[8] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, Y. Wang, and E. Chen, "Large language models for generative information extraction: A survey," 2024.

[9] Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu, "Unified structure generation for universal information extraction," 2022.

[10] O. Sainz, I. García-Ferrero, R. Agerri, O. L. de Lacalle, G. Rigau, and E. Agirre, "Gollie: Annotation guidelines improve zero-shot information-extraction," 2024.

[11] C. R. Fletcher and C. P. Bloom, "Causal reasoning in the comprehension of simple narrative texts," *Journal of Memory and Language*, vol. 27, no. 3, pp. 235–244, 1988.

[12] X. Liu, Z. Luo, and H. Huang, "Jointly multiple events extraction via attention-based graph information aggregation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2018.

[13] S. Yang, D. Feng, L. Qiao, Z. Kan, and D. Li, "Exploring pre-trained language models for event extraction and generation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 5284–5294, Association for Computational Linguistics, July 2019.

[14] J. Yang, S. C. Han, and J. Poon, "A survey on extraction of causal relations from natural language text," 2021.

[15] W. Ali, W. Zuo, W. Ying, R. Ali, G. Rahman, and I. Ullah, "Causality extraction: A comprehensive survey and new perspective," *College of Computer Science and Technology, Jilin University, China*, 2022.

[16] K. Liu, Y. Chen, J. Liu, X. Zuo, and J. Zhao, "Extracting events and their relations from texts: A survey on recent research progress and challenges," *AI Open*, vol. 2, pp. 43–50, 2021.

[17] S. Zhao, M. Hu, Z. Cai, and F. Liu, "Modeling dense cross-modal interactions for joint entity-relation extraction," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (C. Bessiere, ed.), pp. 4032–4038, International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[18] G. Bekoulis, J. Deleu, T. Demeester, and C. Develder, "Joint entity recognition and relation extraction as a multi-head selection problem," *Expert Systems with Applications*, vol. 114, p. 34–45, Dec. 2018.

[19] M. Eberts *et al.*, "Span-based joint entity and relation extraction with transformer pre-training," in *proceedings of ECAI 2020*, 2020.

[20] U. Zaratiana, N. Tomeh, N. E. Khbir, P. Holat, and T. Charnois, "Grapher: A structure-aware text-to-graph model for entity and relation extraction," 2024.

[21] J. Kim, D. T. Nguyen, S. Min, S. Cho, M. Lee, H. Lee, and S. Hong, "Pure transformers are powerful graph learners," in *Advances in Neural Information Processing Systems* (A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds.), 2022.

[22] B. Paassen, D. Grattarola, D. Zambon, C. Alippi, and B. E. Hammer, "Graph edit networks," in *International Conference on Learning Representations*, 2021.

[23] D. Ye, Y. Lin, P. Li, and M. Sun, "Packed levitated marker for entity and relation extraction," 2022.

[24] Z. Zhong and D. Chen, "A frustratingly easy approach for entity and relation extraction," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, eds.), (Online), pp. 50–61, Association for Computational Linguistics, June 2021.

[25] Y. Zhao, X. Jin, Y. Wang, and X. Cheng, "Document embedding enhanced event detection with hierarchical and supervised attention," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Melbourne, Australia), pp. 414–419, Association for Computational Linguistics, July 2018.

[26] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," 2019.

[27] Y. Luan, D. Wadden, L. He, A. Shah, M. Ostendorf, and H. Hajishirzi, "A general framework for information extraction using dynamic span graphs," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 3036–3046, Association for Computational Linguistics, June 2019.

[28] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, J. Kang, J. Yang, S. Li, and C. Du, "Instructuie: Multi-task instruction tuning for unified information extraction," 2023.

[29] C. Kruengkrai, K. Torisawa, C. Hashimoto, J. Kloetzer, J.-H. Oh, and M. Tanaka, "Improving event causality recognition with multiple background knowledge sources using

multi-column convolutional neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, Feb. 2017.

[30] E. Martínez-Cámara, V. Shwartz, I. Gurevych, and I. Dagan, "Neural disambiguation of causal lexical markers based on context," in *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers* (C. Gardent and C. Retoré, eds.), 2017.

[31] S. Liang, W. Zuo, Z. Shi, S. Wang, J. Wang, and X. Zuo, "A multi-level neural network for implicit causality detection in web texts," 2021.

[32] J. Liu, Y. Chen, and J. Zhao, "Knowledge enhanced event causality identification with mention masking generalizations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (C. Bessiere, ed.), pp. 3608–3614, International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.

[33] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020.

[34] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," 2021.

[35] S. K. Sahu, F. Christopoulou, M. Miwa, and S. Ananiadou, "Inter-sentence relation extraction with document-level graph convolutional neural network," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (A. Korhonen, D. Traum, and L. Màrquez, eds.), (Florence, Italy), pp. 4309–4316, Association for Computational Linguistics, July 2019.

[36] C. Yuan, H. Huang, Y. Cao, and Y. Wen, "Discriminative reasoning with sparse event representation for document-level event-event relation extraction," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 16222–16234, Association for Computational Linguistics, July 2023.

[37] Google, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024.

[38] J. Li, M. Wang, Z. Zheng, and M. Zhang, "LooGLE: Can long-context language models understand long contexts?," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (L.-W. Ku, A. Martins, and

V. Srikumar, eds.), (Bangkok, Thailand), pp. 16304–16333, Association for Computational Linguistics, Aug. 2024.

[39] O. C. User, "Reasoning degradation in llms with long context windows." https://community.openai.com/t/reasoning-degradation-in-llms-with-long-context-windows-new-benchmarks/906891/5, 2024. Accessed: 2024-09-13.

[40] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[41] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, 2020.

[42] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "Flair: An easy-to-use framework for state-of-the-art nlp," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, Association for Computational Linguistics, 2019.

[43] B. Ji, S. Li, H. Xu, J. Yu, J. Ma, H. Liu, and J. Yang, "Span-based joint entity and relation extraction augmented with sequence tagging mechanism," *arXiv preprint arXiv:2210.12720*, 2022.

[44] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2524–2527, IEEE, 1997.

[45] U. Zaratiana, P. Holat, N. Tomeh, and T. Charnois, "Hierarchical transformer model for scientific named entity recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[46] F. A. Tan, X. Zuo, and S. Ng, "Unicausal: Unified benchmark and repository for causal text mining," in *Big Data Analytics and Knowledge Discovery - 25th International Conference, DaWaK 2023, Penang, Malaysia, August 28-30, 2023, Proceedings* (R. Wrembel, J. Gamper, G. Kotsis, A. M. Tjoa, and I. Khalil, eds.), vol. 14148 of *Lecture Notes in Computer Science*, pp. 248–262, Springer, 2023.

[47] D. Mariko, H. Abi-Akl, K. Trottier, and M. El-Haj, "The financial causality extraction shared task (fincausal 2022)," in *Proceedings of the 4th Financial Narrative Processing*

Workshop @LREC2022, (Marseille, France), pp. 105–107, European Language Resources Association, June 2022.

[48] R. Prasad, N. Dinesh, A. Lee, A. Joshi, and B. Webber, "Attribution and its annotation in the penn discourse treebank," in *Proceedings of the Linguistic Annotation Workshop*, (Prague, Czech Republic), pp. 31–38, Association for Computational Linguistics, 2007.

[49] R. Prasad, B. Webber, A. Lee, and A. Joshi, "Penn discourse treebank version 3.0," 2019.

[50] C. Hidey and K. McKeown, "Identifying causal relations using parallel wikipedia articles," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 1424–1433, Association for Computational Linguistics, 2016.

[51] J. Dunietz, L. Levin, and J. Carbonell, "The because corpus 2.0: Annotating causality and overlapping relations," in *Proceedings of the 11th Linguistic Annotation Workshop*, (Valencia, Spain), pp. 95–104, Association for Computational Linguistics, 2017.

[52] P. Mirza, R. Sprugnoli, S. Tonelli, and M. Speranza, "Annotating causality in the TempEval-3 corpus," in *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)* (O. Kolomiyets, M.-F. Moens, M. Palmer, J. Pustejovsky, and S. Bethard, eds.), (Gothenburg, Sweden), pp. 10–19, Association for Computational Linguistics, Apr. 2014.

[53] T. Caselli and P. Vossen, "Eventstoryline: Creating a cross-lingual event-centric timeline," in *Proceedings of the 1st Workshop on Computing News Storylines (CNS 2017)*, (Vancouver, Canada), pp. 1–11, Association for Computational Linguistics, 2017.

[54] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. O'Seaghdha, S. Pado, M. Pennacchiotti, L. Romano, and S. Szpakowicz, "Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, (Uppsala, Sweden), pp. 33–38, Association for Computational Linguistics, 2010.

[55] C. Hidey and K. McKeown, "Identifying causal relations using parallel wikipedia articles," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1424–1433, Association for Computational Linguistics, 2016.

[56] X. Wang, Z. Wang, X. Han, W. Jiang, R. Han, Z. Liu, J. Li, P. Li, Y. Lin, and J. Zhou, "Maven: A massive general domain event detection dataset," in *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1652–1671, Association for Computational Linguistics, 2020.

[57] T. Satyapanich, F. Ferraro, and T. Finin, "Casie: Extracting cybersecurity event information from text," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 8749–8757, 2020.

[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations (ICLR)*, 2019.

[59] R. Zhang, Y. Li, Y. Ma, M. Zhou, and L. Zou, "Llmaaa: Making large language models as active annotators," 2023.