DocScript: Document-Level Script Event Prediction

Puneet Mathur[*1], **Rajiv Jain**[2], **Vlad Morariu**[2], **Aparna Garimella**[2],
Franck Dernoncourt[2], **Jiuxiang Gu**[2], **Ramit Sawhney**[3], **Preslav Nakov**[3], **and Dinesh Manocha**[1]
[1] University of Maryland, College Park
[2] Adobe Research
[3] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
* Corresponding Author: puneetm@umd.edu

Abstract

We present a novel task of document-level script event prediction, which aims to predict the next event given a candidate list of narrative events in long-form documents. To enable this, we introduce DocSEP, a challenging dataset in two new domains - contractual documents and Wikipedia articles, where timeline events may be paragraphs apart and may require multi-hop temporal and causal reasoning. We benchmark existing baselines and present a novel architecture called DocScript to learn sequential ordering between events at the document scale. Our experimental results on the DocSEP dataset demonstrate that learning longer-range dependencies between events is a key challenge and show that contemporary LLMs such as ChatGPT and FlanT5 struggle to solve this task, indicating their lack of reasoning abilities for understanding causal relationships and temporal sequences within long texts.

Keywords: information extraction, document-level extraction, script event prediction

1. Introduction

Understanding event semantics and their associated relations is a crucial task in natural language processing, with applications in discourse understanding (Nie et al., 2019), automated storyline extraction (Swanson and Gordon, 2008), and inference of missing events (Zhou et al., 2022b). One of the most challenging tasks in this domain is script event prediction (Chambers and Jurafsky, 2008), which aims to predict the subsequent event from a candidate list following a chain of given events.

Prior work has focused extensively on script event reasoning at the sentence level, mostly ignoring events at the document level spanning multiple paragraphs or the full document. Real-world documents such as contracts and articles may be long-range with multiple pages and hundreds of paragraphs. Such documents may have information about related events located far apart. Extracting the sequence of event occurrences is a non-trivial problem requiring multi-hop reasoning and contextual understanding of event temporality and causality. Current evaluation datasets for script event prediction, such as MCNC (Granroth and Clark, 2016a), exhibit a low prevalence of global discourse-level event chain annotations, wherein local syntactic cues (Wang et al., 2017a; Lv et al., 2019) and sentence-level semantic understanding (Li et al., 2018; Bai et al., 2021b) helps achieve near human-level performance. In this work, we extend the script event prediction task to document scale to propose the Document-level Script Event Prediction task (see Figure 2), which requires discourse-level contextual understanding as well as multi-hop reasoning to predict the logical order of event occurrence in the text.

To address these long-standing challenges, we introduce **DocSEP - Document-level Script Event Prediction**, a novel dataset for document-level script event prediction in two new domains: contractual documents and Wikipedia articles. Contract documents contain rich event timelines and cause-effect conditions. Understanding the logical sequence of events in contracts is an important business problem with legal and monetary consequences. Wikipedia articles serve as important sources of rich knowledge semantics containing diverse event information with relational facts. The context and the candidate events are extracted from long multi-page contracts and Wikipedia articles ($\sim$ 2500 words). Events may be located several sentences or paragraphs apart and require multi-hop temporal and causal reasoning.

Document-level script event prediction requires the model to consider global event-event interactions to resolve temporal and causal relations between events to recover the chain of logical event sequences. Prior work has looked at using local semantics or heuristics-based graph learning techniques. Past research has explored the use of local event co-occurrence graphs (Li et al., 2018; Zheng et al., 2020a; Lee et al., 2020), but did not explicitly utilize sequential event logic graphs from narrative texts, which can serve as a source of external structured knowledge for event correlation. Recent generative LLMs like FlanT5-XXL (Chung et al., 2022), LlaMA (Touvron et al., 2023), and ChatGPT have exhibited remarkable ability in zero-shot learning yet struggle to outperform supervised models

LREC-COLING 2024, pages 5140–5155
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

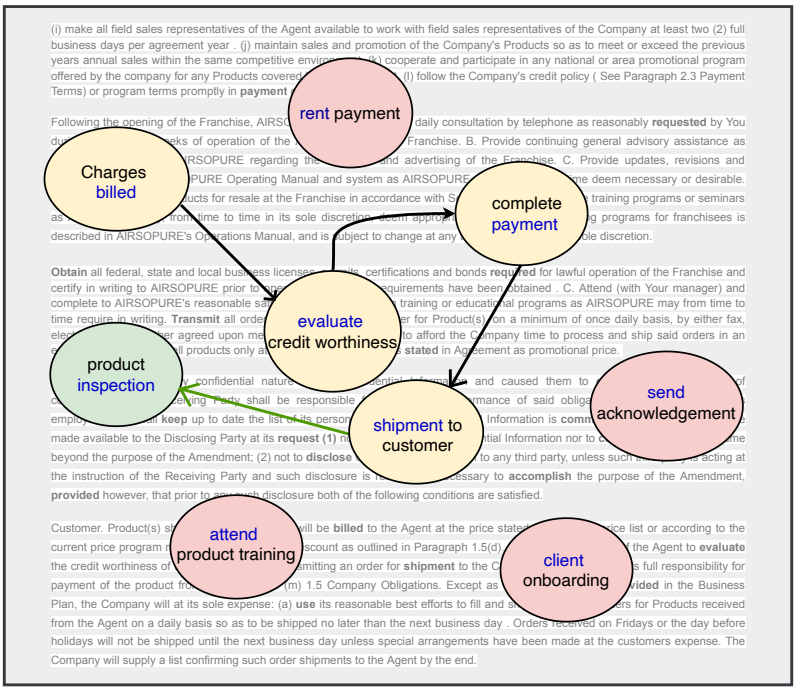


Figure 1: The document-level script event prediction task involves choosing the most reasonable script event (green) following a chain of discourse-level context events (yellow) from a set of candidate events (red) spanning multiple paragraphs in a long text document.

for long-dependency temporal relation extraction and script event prediction (Yuan et al., 2023). The temporal and causal relations predicted by LLMs or task-specific supervised models can be noisy and unreliable due to a lack of high-quality labeled data and hallucinations in textual grounding. We hypothesize that these mispredictions can be corrected by optimizing the model to find an alignment of event-event relations that satisfy semantic constraints with the least amount of penalty incurred due to logical constraints. Optimal transport is an established mechanism to efficiently find an alignment plan between two groups of points (i.e., global event–event relation label distribution in our case) based on their pairwise transportation costs and the distribution mass accumulated on the points. We hypothesize that the optimal transport-based optimization approach can explicitly recover underlying global temporal and causal relations throughout the document and aid in predicting the next event following a chain of script events.

In the era of LLMs and prompt-based generation, script event prediction still holds major significance for understanding event-centric discourse in long context documents. LLMs have been shown to struggle in grounding event understanding to document context due to their strong tendency to hallucinate. Supervised script event prediction methods can provide a reliable alternative for exploiting temporal and causal reasoning for event chain extraction. We introduce `DocScript`, a novel script event prediction model to efficiently find an optimal alignment plan between noisy and target event pair relations to learn the logical order of event sequences in long-context narrative text. Further, most pre-trained language models (PLM), such as BERT, T5, and GPT, use token-level learning objectives and thus are unable to capture the correlations between events, ignoring event-specific knowledge. To better augment the generalized knowledge of pre-trained models with event knowledge for script event prediction tasks, we propose novel event-aware instruction-tuning tasks for the base Transformer language model - Event Chain Coherence Detection, Re-ordered Event Sequence Generation, and Event Pair Connectivity Identification, to effectively learn sequential event reasoning. Our work offers two main contributions:

1. We propose DocSEP - Document-level Script Event Prediction, a novel dataset for document-level script event prediction in two new domains: contractual documents and Wikipedia articles.

2. We benchmark several baseline models on the document level script event prediction task to study their ability to learn the logical order of event sequences in long-context narrative text and effectiveness in utilizing temporal and causal reasoning for subsequent event prediction.

2. Related Work

Script Event Prediction: Previous studies have used pair-wise relationships (Chambers and Jurafsky, 2008), composite word embedding functions (Granroth and Clark, 2016a), temporal ordering (Jans et al., 2012a; Pichotta and Mooney, 2016), and event chain information (Wang et al., 2017b; Lv et al., 2019) for predicting the subsequent event. Prior work has also explored statistical correlations between event nodes in graph structures (Li et al., 2018; Zheng et al., 2020a; Lee et al., 2020) and attention (Wilner et al., 2021). A recent line of work combined the strength of transformer-based models with graph-structured data. (Zheng et al., 2020b; Du et al., 2022a) encoded knowledge graphs using graph embedding to integrate entity representation with language models. (Guan et al., 2018) proposed a retrieval-based method to leverage structural information of knowledge graphs. However, constructing complete event logic graphs that retrieve related context events is non-trivial and error-prone due to noisy temporal and causal dependency extraction techniques (Mathur et al., 2022; Chen et al., 2022). Contemporary event understanding methods utilize pre-trained LMs such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019a) that lack discourse-aware event correlations as they are trained via Masked Language Modeling, which does not capture causal and temporal relations between multi-hop events required for document-level script event prediction.

3. Document-level Script Event Prediction Task Formulation

Let document D be a sequence of n tokens $[x_1, \cdots, x_n]$ and p events (e_{i1}^{p}). For a given event chain of $m-1$ lying on an event chain (referred to as context events) $C = (e_1, e_2, ..., e_{m-1})$, the document-level script event prediction chooses the most suitable subsequent event e_m from a candidate list $O = \{o_1, o_2, \cdots, o_l\}$. The set of candidate events is sampled from the superset of document events except the ones identified as context events ($O \in \{e_i\}_1^p - C$).

4. DocSEP Datasets

Next, we propose DocSEP dataset - the first corpus that explicitly focuses on event chains spanning several sentences and paragraphs in a document. Table 3 emphasizes the key differences between the existing MCNC and the proposed DocSEP dataset. DocSEP has two sub-variants: (1) DocSEP-Contracts - derived from legal contracts (2) DocSEP-MavenWiki comprising of Wikipedia articles.

4.1. Data Acquisition

The DocSEP-Contracts corpus is formed by extracting 100 contractual documents from the Atticus dataset[1] (Hendrycks et al., 2021a), which were originally sourced from public domain SEC contracts. Due to the multi-page length of these documents, we limited the annotations to the first 2500 words, excluding the definition sections since they did not contain many events for this task. We construct DocSEP-MavenWiki by utilizing Wikipedia articles from the MAVEN-ERE (Wang et al., 2022) corpus.

4.2. Constructing Event Logic Graph

Event Logic Graphs (ELGs) are directed cyclic graphs whose nodes are events and whose edges stand for the temporal, causal, or conditional relations between events. Essentially, ELGs are an event-centric knowledge base that reveals evolutionary patterns and connected logic of real-world events (Ding et al., 2019). We construct an ELG graph corresponding to each document, which is then utilized to extract script event sequences at the discourse level. To obtain the ELG annotations required for our task, we follow (Ding et al., 2019) and perform the following steps - (1) Event and Time Expression (TIMEX) Span Extraction, (2) Temporal Relation Extraction and Dependency Parsing, (3) Conditionality and Causality Extraction. Wikipedia documents for DocSEP-MavenWiki sourced from MAVEN-ERE already have annotations for Task (1)-(3). We obtain manual annotations for documents required for DocSEP-Contracts corpus.

Following annotation, we perform Sequential Relation Parsing to select all sequential relations between event nodes. We resolve all relations to follow forward logic from tail to head event. To achieve this, we reverse Before relations to form After and align conditional relations from condition to action. We merge subsets of Overlap and Equal relations into single event nodes. For all annotation, we use the BRAT tool[2] (Stenetorp et al., 2012) and employ 5 freelance annotators from Upwork with past experience in linguistic annotations due to the difficulty of legal language. Annotators are provided with example annotations and guidelines for each labeling task. We train two sets of annotators per task if they pass a qualifying task to label a small set of 100 annotations with high accuracy to increase the dataset quality. All annotations were verified by an expert annotator with a background in computational linguistics to resolve conflicts. While our focus for this work is script

[1] https://www.atticusprojectai.org/cuad
[2] https://brat.nlplab.org/

Feature	DocSEP-Contract	DocSEP-MavenWiki
Events	0.79	0.76
Timex	0.81	0.83
Temporal Dependency Links	0.62	0.67
Causal Dependency Links	0.64	0.69
Script Event Prediction	0.76	0.79

Table 1: Inter-annotator agreement for DocSEP dataset using Cohen's Kappa.

Statistics	MCNC	DocSEP-Contracts	DocSEP-Wiki
Documents	1.03M	98	2310
Events	522K	27K	49K
Timex	-	1628	12046
Temporal Relations	-	20K	316K
Causal Relations	-	1874	28K
Script Event (SE) Chains	160K	128K	5.64M
SE with Temporal Connective only	-	126K	5.28M
SE with Causal Connective only	-	1681	356K
Average Document Length	31	2479.23	232.79
Average Sequence Word Distance	8	212.66	174.50

Table 2: Data Statistics comparing MCNC (Granroth and Clark, 2016a) with DocSEP Contracts and Wiki.

event prediction, we believe that the event-logic graph developed here for DocSEP-Contracts may be useful for the community working on legal NLP.

4.3. Script Event Sequence Generation

In order to extract script event sequences needed for our task, we perform Depth First Search on the Event Logic Graph to select all event chains of lengths greater than four, ignoring the TIMEX nodes. We extract only 4 context events per data sample following the MCNC dataset. Our methodology can be extended for longer event chains as well which we defer to future work. For each document, we sample subsets of four consecutive events from the event chain between the start and end node such that all event pairs are separated by multiple sentences/paragraphs. Each instance in the DocSEP dataset is paired with five options, out of which only one is the correct ground truth, and the remaining is incorrect. To create multiple choices, random events not belonging to the same sub-chain are sampled from the document while ensuring the event options are all sampled from a threshold neighborhood of 1000 words above or below the nearest context event to the correct event option in the text document. The incorrect events were carefully chosen such that they do not share any common ancestors in the graph. Consequently, this ensured that the "incorrect options" in MCQ were always non-related to the context events and diminished the possibility of cases where multiple options are correct script events.

4.4. Data Quality Estimate

Table 1 reports Cohen's kappa for the ELG construction in DocSEP. We report a high degree of inter-annotator agreement across all sub-tasks and the main script event prediction task. For DocSEP-MavenWiki, IAA metrics for ELG are sourced from (Wang et al., 2022)). We hired two experienced linguistic annotators to manually annotate and verify the correctness of script event MCQs in the evaluation split for each dataset. Annotators were asked to choose the most plausible subsequent event for each MCQ. In cases of conflict, a linguistic expert was asked to select the ground truth annotation. We report IAA for this task in Table 1 and as a human benchmark evaluation in Table 4 on the full test set. We observed a high accuracy of 87.33% and 91.45%, respectively, enhancing our confidence in the correctness of the MCQ setup. Note only 22.1% of the MCQs have their sequence of the chain events to be the same as their textual order, which is close to the random answer selection baseline (20%). Annotators were also asked to mark if more than one option can be considered the correct next script event to assess the dataset ambiguity. We found that less than 5% of MCQ samples had multiple possible correct options (4.6% for DocSEP-Contracts and 3.8% for DocSEP-MavenWiki), which were then discarded.

Data Statistics: We create a train, validation, and test split for both DocSEP-Contracts and DocSEP-MavenWiki with a ratio of 70:10:20. Each split has unique documents to prevent data leakage. Table 2 describes the data statistics comparing MCNC (Granroth and Clark, 2016a) dataset with

Statistic	MCNC	DocSEP-Contracts	DocSEP-MavenWiki
Avg Doc Length	Sentence-level (∼ 31)	Multiple Pages (∼ 2480)	Multiple Paragraphs (∼ 250)
Temporal Logic	✗	✓	✓
Causal Logic	✗	✓	✓
Domain	News	Legal Contracts	Wikipedia Articles

Table 3: Comparison of DocSep with Multi-Choice Narrative Cloze (MCNC) dataset (Granroth and Clark, 2016a). DocSEP is the largest discourse corpus for script event prediction tasks on contracts and Wikipedia articles.

DocSEP corpus. MCNC has a comparable number of event chains to DocSEP-Contracts, while DocSEP-MavenWiki has 50x more script event annotations. Overall, DocSEP has a significantly higher average sequence word distance and document length measured in the number of words, allowing investigation of document-level script event prediction.

5. DocScript

We introduce DocScript (see Figure 2)- a novel neural architecture for document-level script event prediction that uses Optimal Transport between event pair representations and edges of noisy event logic graphs extracted from the input document. It leverages three unique event-aware instruction fine-tuning tasks to help the model learn the correct order of logical event sequences.

5.1. Document Encoding

We prepend the special tokens <e> in front of each event in the input document D, except for context and candidate events, which are prefixed by <c_i> and <o_i>, respectively. We leverage LongT5 - a long-context pre-trained transformer to obtain the embedding of each word in the document. The output embeddings (last hidden states) from the LongT5 encoder and the decoder are represented as $H_E = LongT5\_Enc([x_1, x_2, < e >, ..., x_n])$ and $H_D = LongT5\_Dec([x_1, x_2, < e >, ..., x_n])$.

5.2. Event-aware Instruction Tuning

Inspired by Wei et al. (2021), we perform event-aware instruction fine-tuning of the LongT5 model. **1. Event Chain Coherence Detection**: We want the model to be able to recognize a correctly ordered event sequence from a scrambled one. Given a document D in the train set, an event chain $s = (e_1, e_2, e_3, \cdots, e_r)$ is extracted. We augment each positive example with a negative event sequence by scrambling the tagged special tokens for, e.g., $s^- = (e_1, e_3, e_2, \cdots, e_r)$. We construct a prompt that incorporates the text document followed by one of the positive or negative event sequences (s^+ or s^-). We add the suffix instruction - "*Is the sequence of events in the correct order? Answer in yes or no only.*"

2. Re-ordered Event Sequence Generation involves prompting the LongT5 model to generate correctly ordered event sequences. Given a document D, an event chain is extracted from its Event Logic Graph as $s = (e_1, e_2, e_3, \cdots, e_f, \cdots, e_r)$. We generate a negative sample by swapping out one of the events e_f with another event $e_f^{'}$ not lying on the event chain to form $s^{'} = (e_1, e_2, e_3, \cdots, e_f^{'}, \cdots, e_r)$. For the event to be replaced, we change the position of the corresponding special token $< e_i >$ to the false event. The document text is concatenated with the scrambled event sequence. We then append the suffix instruction -"*The correct order of the event sequence is*," and we pass it through the LongT5 model, which is supposed to generate the correct sequence of events.

3. Event Pair Connectivity Identification Given two events in the document, we want the model to learn to predict if the events lie on the same event chain in the logic graph. For two selected events e_i and e_j in the document text, we construct a prompt with document text followed by a suffix: "*Are events e_i and e_j on the same event chain? Answer true or false only*". We expect the model to generate *true/false* as the answer during prompt training of the LongT5 model, which we then compare to the ground truth text.

5.3. Event-Pair Representation Learning

We build an event-pair relation matrix $F \in R^{p^2 \times d}$ to capture the correlations between event pairs (e_i, e_j). We use the embeddings corresponding to the special tokens in LongT5 encoder output to compute an event-pair feature vector $F(e_i, e_j)$ by concatenating the element-wise similarity ($\odot$), the cosine similarity ($cos(,)$), and the bi-linear similarity between h_{E_i} and h_{E_j} as: $F(e_i, e_j) = [h_{E_i} \odot h_{E_j}; cos(h_{E_i}, h_{E_j}); h_{E_i} W_1 h_{E_j}]$. We hypothesize that global interactions between event pairs can help learn interdependencies between discourse-level chain events and subsequently guide the script event prediction task. Inspired by (Zhang et al., 2021), we use U-Net (Ronneberger et al., 2015) architecture to enhance information exchange between event pairs at the document level for implicit reasoning. The U-net architecture U contains two down-sampling and two up-sampling

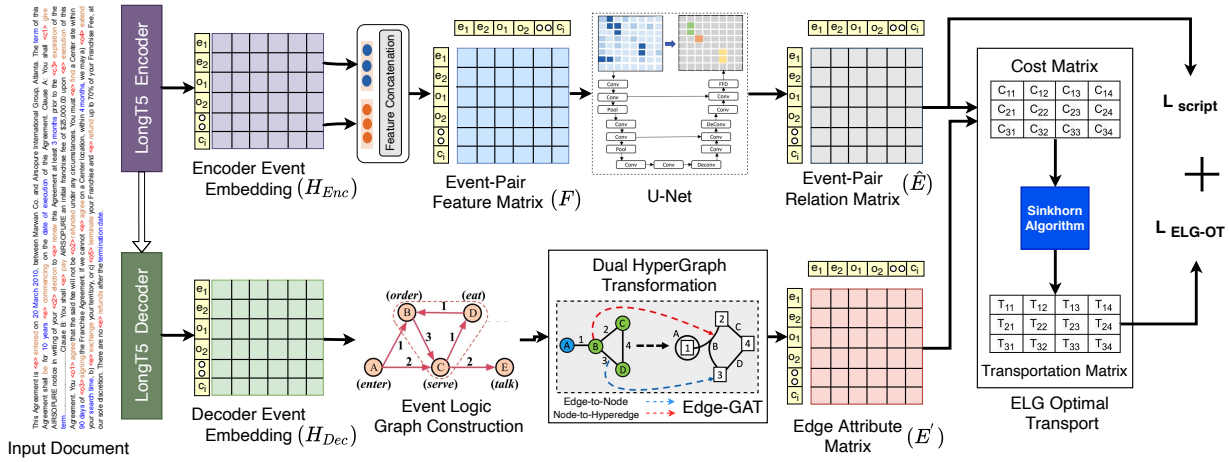


Figure 2: DocScript Model: The input document is passed through a LongT5 transformer model instruction-tuned with event-aware self-supervised tasks to obtain hidden states from the encoder (H_{Enc}) and the decoder (H_{Dec}). The event embeddings from the encoder are pairwise combined to form an Event-pair Feature Matrix, which is then passed through U-Net architecture to obtain the Event-Pair Relation Matrix ($\hat{E}$). A weakly-labeled Event Logic Graph (ELG) is built by extracting noisy temporal and causal relations between event nodes and is transformed into its equivalent hypergraph to convert the edges to nodes and the nodes to hyper-edges using Dual Hypergraph Transform (DHT). The hypergraph is passed through the Edge-GAT network to extract the edge attributes corresponding to each node pair. We perform Optimal Transport between the Event-Pair Relation Matrix $\hat{E}$ and Edge Attribute Matrix $E^{'}$ to optimize the transportation matrix with a cosine function-based cost matrix. We combine the ELG-Optimal Transport loss L_{ELG-OT} with Script Event Prediction loss L_{script} for the script event prediction task.

blocks with skip connections to capture both local and global information as $\hat{E} = U(W_2 F)$. W_1, W_2 are learnable weights.

5.4. Script Event Prediction as Event Logic Graph Optimal Transport

Recovering the logical interdependencies between event pairs can guide script event prediction at the discourse level. Hence, we train the DocScript model to match the learned semantic representations of event pairs in the relation matrix with the corresponding edge features extracted from the document event logic graph via Optimal Transport.

Document Event Logic Graphs (ELG): We extract the Event Logic Graph (ELG) from each input document D as a directed cyclic graph G, where the vertices are event mentions e_i connected by edges that represent temporal t or causal c relations between event pairs. The ELG node embeddings are initialized with the special event tokens embeddings from the decoder output H_{Dec}. To extract the contextual edge representation from the ELG, we exploit hypergraph duality that transforms the original ELG into an equivalent hypergraph with the ELG edges and nodes mapped to respective nodes and hyperedges of the hypergraph. The hypothesis behind this transformation is that if we can change the role of the nodes and of the edges of the graph with a shared connectivity pattern while accurately preserving their information, we can use node-based message-passing schemes in hypergraph to learn the ELG edge representations. Let the Event-Logic Graph $G = (V, A, E)$ be defined by its node feature matrix $V \in R^{p \times d}$, incidence matrix $A \in \{0, 1\}^{p \times q}$, and initial edge attribute $E \in R^{q \times d}$ for q edges each of which is a one-hot encoding of relation types. The incidence matrix A denotes interactions between nodes and edges. We perform Dual Hypergraph Transformation (DHT) (Jo et al., 2021) to interchange the node and the edge features of the original ELG into a hypergraph followed by edge-based GAT convolution message-passing (EGAT) (Chen and Chen, 2021) between nodes of the dual hypergraph to effectively represent the edge attributes of the original Event-Logic graph.

We now describe the approach to utilize Event-Logic Graph Optimal Transport (ELG-OT) for document-level script event prediction. We treat the event pair feature matrix $\hat{E}$ and edge attribute matrix from the document ELG $E^{'}$ as the source and target matrices. Optimal transport aims to compute a minimal cost transportation plan between a source distribution μ_s, and a target distribution μ_t defined on discrete probability spaces $X, Y \in \Omega$, respectively. We compute the ELG optimal transport loss using the Sinkhorn-Knopp algorithm (Cuturi, 2013) using cosine distance as the cost function: $L_{ELG-OT} = Sinkhorn(\hat{E}, E^{'})$.

Script Event Prediction: For all given script event candidates in an event chain, we perform softmax over their learned event pair feature $\hat{E}$ between the candidate event o_i and last context event e_{m-1} to get the final scores as $P(o_i|e_1, e_2, \cdots, e_{m-1}) =$

$\frac{\exp(\hat{E}_{i,m-1})}{\sum_j \exp(\hat{E}_{j,m-1})}$. During training, we select the candidate event with maximum probability as the predicted event: $\arg\max_i P(o_i|e_1, e_2, \cdots, e_{m-1})$. Given an event chain and a set of event candidates, our goal is to minimize the cross-entropy loss between correct answers and predictions as

$$L_{script} = -\frac{1}{N}\sum_{k=1}^{N}\sum_{i}\log(P(o_i|e_1, \cdots, e_{m-1}))$$

5.5. Training DocScript

Multi-task Learning: We simultaneously optimize script event prediction loss (L_{script}) and ELG-Optimal Transport loss (L_{ELG-OT}) as they both reinforce each other owing to their shared event semantics. The final optimization takes their weighted sum as $L = \lambda L_{script} + (1-\lambda)L_{ELG-OT}$, where λ is a hyperparameter.

6. Experiments

Baselines: We compare the following models: **Event Pair and Event Chain Based Methods**: **(i) PMI** (Chambers and Jurafsky, 2008) is the co-occurrence-based model which adopts Pointwise Mutual Information (PMI) to score candidate events and it does not use the entire text, **(ii) Bigram** (Jans et al., 2012b) computes event pair relations based on bigram probabilities. **(iii) Word2vec** (Mikolov et al., 2013) learns word embeddings of events from training text corpora, **(iv) Event-Comp** (Granroth and Clark, 2016b) learns the strength of association between two events using a coherence function, **(v) PairLSTM** (Wang et al., 2017a) integrates event order information using LSTM, **(vi) Sam-Net** (Lv et al., 2019) integrates event-level and chain-level attentions through adopts Dense-Net. **Event Graph-Based Methods**: **(i) SGNN** (Li et al., 2018) learns a narrative event evolutionary graph (NEEG) using GNNs, **(ii) HeterEvent** (Zheng et al., 2020a) encodes word-event relationship by graph attention mechanism as input to BERT. **Pretrained Language Model Based Methods**: **(i) BERT** (Devlin et al., 2019) fine-tuned on truncated documents containing the context and option events. **(ii) LongT5** (Guo et al., 2022) is a long-range Transformer models for long documents. We also compare recent promptable LLMs like ChatGPT (GPT-3.5 Turbo) and FlanT5-XXL (Chung et al., 2022) by utilizing two types of prompt strategies - (1) Zero-shot: Given the document, we design a prompt with full document text followed by pairwise reasoning of consecutive context events in the prompt (e.g. "E_j is subsequent script event to E_i"), and the query - "Given the text and chain of events, choose the most suitable next subsequent event? "; (2) Chain-of-Thought Prompting: We design a two-stage CoT prompt where we first ask the model to determine the temporal/causal relation between the option event and the last context event in the chain followed by prompting for options for which the model predicts "after", "overlaps", or "causal" relations. More details in the Appendix.

Training Setup - Dataset and Metrics: We train and evaluate on DocSEP-Contracts and DocSEP-MavenWiki from Section-4. Following prior work of (Granroth and Clark, 2016a; Li et al., 2018), we utilize the accuracy (%) of choosing the correct subsequent event from five multiple-choice options as the evaluation metric.

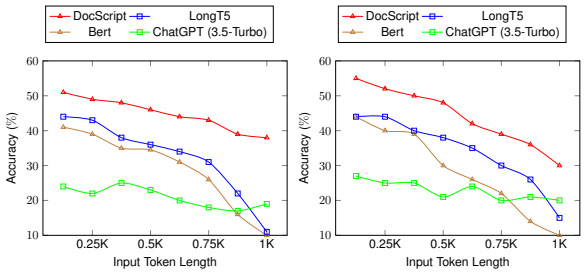
7. Results

Overall Performance Comparison: Table 4 compares the performance all methods on DocSEP-Contracts and DocSEP-MavenWiki datasets. We observe that event pair models such as PMI (Chambers and Jurafsky, 2008), Bigram (Jans et al., 2012b), Pair-LSTM (Wang et al., 2017b), and EventComp (Granroth and Clark, 2016a) perform poorly on DocSEP, much lower than the majority baseline. Sam-Net (Lv et al., 2019), SGNN(Li et al., 2018), and HeterEvent (Zheng et al., 2020b) leverage sentence-level event chains and graph structures yet struggle due to distractor interactions confounding the reasoning process in densely-connected graphs. Methods utilizing Transformer language models like BERT (Devlin et al., 2018) are challenged by their input length restriction of 512 tokens for contextual reasoning over longer input lengths. Transformer models such as LongT5 (Guo et al., 2022) for supervised classification over candidate event mentions. They both perform marginally better than the BERT but are challenged due to their inability to reason in a multi-hop fashion. We observe that DocScript is better than the generative BART model (Zhu et al., 2023) and outperforms LongT5 and other Transformer-based baselines by a significant margin DocSEP-Contracts and DocSEP-MavenWiki. Overall, contemporary baseline methods show lower performance on DocSEP due to longer input context, high domain-specificity of contracts and Wikipedia articles, and lack of document-level context for event embeddings. More on ablation results are in the Appendix.

Comparison with Large Language Models: We observe that generative LLMs like ChatGPT and FlanT5-XXL are challenged for script event prediction tasks in the zero-shot setting. FlanT5-XXL generates random words from the document instead of event triggers for prompts exceeding 512 tokens. GPT-3.5 is narrowly better at temporal understanding but fails to infer discourse-level causal

	System	DocSEP-Contracts Acc (%)	DocSEP-MavenWiki Acc (%)
Supervised Baselines	Majority	20.00	20.00
	PMI (Chambers and Jurafsky, 2008)	9.08	7.84
	Bigram (Jans et al., 2012b)	6.67	12.85
	Word2vec (Mikolov et al., 2013)	22.46	26.34
	Event-Comp (Granroth and Clark, 2016b)	22.56	26.85
	PairLSTM (Wang et al., 2017a)	19.36	25.45
	Sam-Net (Lv et al., 2019)	20.73	19.94
	SGNN (Li et al., 2018)	19.84	20.03
	HeterEvent (Zheng et al., 2020a)	25.66	29.35
	BERT (Devlin et al., 2019)	31.25	35.47
	LongT5 (Guo et al., 2022)	34.27	38.87
LLMs	FlanT5-XXL (Chung et al., 2022)	16.65	17.95
	GPT-3.5-turbo Zero-shot	19.45	21.36
	ChatGPT CoT	25.50	26.67
	DocScript	39.16	42.55
	Human	87.33	91.45

Table 4: Performance comparison of methods for script event prediction on DocSEP-Contracts and DocSEP-MavenWiki.



(a) DocSEP-Contracts (b) DocSEP-MavenWiki

Figure 3: Plot of accuracy % vs token length of the event chains in input document.

information. ChatGPT (GPT-3.5) with CoT reasoning stands slightly better due to step-wise relation inference but still lags behind Transformer models like BERT. These results echo challenges in LLM temporal reasoning found by (Yuan et al., 2023) and suggest this task remains relevant in the era of LLMs and GPT. Smaller supervised models prove better than zero-shot LLMs in this domain.

Impact of Long-context Event Chains: Figure 3 plots the accuracy vs token length of the event chain in the input document. The performance of the BERT model degrades rapidly as the length increases to over 500 tokens. LongT5 shows improved performance but decays due to the lack of structural knowledge about long-form logical event relations, while ChatGPT remains unaffected by length. DocScript model maintains steady improvement over baseline models with increasing input lengths.

Discussion on Dataset: Our paper is the first to provide a labeled corpus for document-level script event prediction evaluation in two new domains - legal documents and Wikipedia articles. The cost and time spent on document-scale annotation were prohibitively high. We believe that our work will directly be helpful for several downstream applications of event timeline extraction in narrative and contractual documents, where temporal reasoning over events has real-world legal and monetary implications for users.

8. Conclusion and Future Work

In conclusion, we introduce a pioneering task in document-level script event prediction, focused on forecasting the subsequent event from a list of narrative events within lengthy documents. To support this endeavor, we introduce the DocSEP dataset, comprising samples extracted from two distinct domains: contractual documents and Wikipedia articles. These domains present challenges, such as events scattered across paragraphs, and necessitate multi-hop temporal and causal reasoning. Our proposed architecture, DocScript, along with refined baseline models, addresses the need to comprehend sequential event ordering at the document level. Future works may expand current methods to other low-resource domains and exploit generative reasoning methods to extract event sequences. Our current work is limited to English. Extending this work to low-resource settings will require exploiting language-specific tools for event and temporal expression extraction. A hard challenge to solve in adapting this work to newer domains is the annotator expertise gap. We believe that utilizing recently released GPT-4 style models for noisy data annotation can act as a good starting point to bridge this gap for new domains.

9. Ethics Statement

We utilize the publicly available MCNC dataset for the script event prediction task. Additionally, we also curated two datasets on contract documents and Wikipedia articles. We source the contract documents from a publicly available resource - ATTICUS, as we repurpose the document in this dataset for our task. Wikipedia article documents for the DocSEP-MavenWiki dataset were sourced from the MAVEN-ERE corpus, which openly releases this data for research purposes. Our proposed DocSEP corpus provides new annotations and does not violate any privacy, as these documents are already in the public domain. There is no human bias involved in such documents as they are business contracts filed on the SEC website. These documents do not restrict reuse for academic purposes, and any personal information was already redacted before their original release. All documents and our experiments are restricted to the English language. We paid the freelance annotators on an average $25 per hour for the entire annotation task. The annotators were informed about the purpose of the annotation and adequate accommodations were provided to ensure there are no adverse effects such as annotation fatigue. The documents to be annotated do not involve any content that is harmful to any sub-population of the annotators as they are sourced from business fillings and public Wikipedia sources. There was no sensitive data involved in the studies.

Potential Risks: Our models are exploratory and academic in nature and should not be used for real-world legal/contractual/healthcare purposes without extensive investigations into their shortcomings/randomness/biases. **Unhandled Cases**: The current work is limited to the English language and would need suitable tools in other languages to process event logic graphs, temporal and causal relations, and language models. Moreover, our method has been tested on limited domains of Wikipedia text, narrative stories, and contracts. Applications in life-critical scenarios such as healthcare, public safety, and law will need further investigation.

10. References

Mohammed Aldawsari, Adrian Perez, Deya Banisakher, and Mark Finlayson. 2020. Distinguishing between foreground and background events in news. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5171–5180.

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

David Alvarez-Melis and T. Jaakkola. 2018. Gromov-wasserstein alignment of word embedding spaces. In *Conference on Empirical Methods in Natural Language Processing*.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021a. Syntax-bert: Improving pre-trained transformers with syntax trees. *arXiv preprint arXiv:2103.04350*.

Long Bai, Saiping Guan, Jiafeng Guo, Zixuan Li, Xiaolong Jin, and Xueqi Cheng. 2021b. Integrating deep event-level and script-level information for script event prediction. In *Conference on Empirical Methods in Natural Language Processing*.

Song Bai, Feihu Zhang, and Philip HS Torr. 2021c. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637.

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, K. McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. *ArXiv*, abs/2004.04295.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *ArXiv*, abs/2004.05150.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from rst discourse parsing. *arXiv preprint arXiv:1509.01599*.

Taylor Cassidy, Bill McDowell, Nathanel Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. Technical report, Carnegie-Mellon Univ Pittsburgh PA.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014a. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014b. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797.

Jiayuan Chen, Boyu Zhang, Yinfei Xu, and Meng Wang. 2021a. Textrgnn: Residual graph neural networks for text classification. *arXiv preprint arXiv:2112.15060*.

Jun Chen and Haopeng Chen. 2021. Edge-featured graph attention network. *arXiv preprint arXiv:2101.07671*.

Liqun Chen, Guoyin Wang, Chenyang Tao, Dinghan Shen, Pengyu Cheng, Xinyuan Zhang, Wenlin Wang, Yizhe Zhang, and Lawrence Carin. 2019. Improving textual network embedding with global attention via optimal transport. *ArXiv*, abs/1906.01840.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. Ergo: Event relational graph transformer for document-level event causality identification. *arXiv preprint arXiv:2204.07434*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021b. A dataset for answering time-sensitive questions. *ArXiv*, abs/2108.06314.

Yu Chen, Lingfei Wu, and Mohammed Zaki. 2020. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. *Advances in Neural Information Processing Systems*, 33.

Shivang Chopra, Ramit Sawhney, Puneet Mathur, and Rajiv Ratn Shah. 2020. Hindi-english hate speech detection: Author profiling, debiasing, and practical perspectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 386–393.

Arijit Ghosh Chowdhury, Ramit Sawhney, Puneet Mathur, Debanjan Mahata, and Rajiv Ratn Shah. 2019. Speak up, fight back! detection of social media disclosures of sexual harassment. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop*, pages 136–146.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Xiao Ding, Zhongyang Li, Ting Liu, and Kuo Liao. 2019. Elg: an event logic graph. *arXiv preprint arXiv:1907.08015*.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

L. Du, Xiao Ding, Yue rong Zhang, Ting Liu, and Bing Qin. 2022a. A graph enhanced bert model for event prediction. In *Findings*.

X. Du, Zixuan Zhang, Sha Li, Ziqi Wang, Pengfei Yu, Hongwei Wang, T. Lai, Xudong Lin, Iris Liu, Ben Zhou, Haoyang Wen, Manling Li, Darryl Hannan, Jie Lei, Hyounghun Kim, Rotem Dror, Haoyu Wang, Michael Regan, Qi Zeng, QING LYU, Charles Yu, Carl N. Edwards, Xiaomeng Jin, Yizhu Jiao, Ghazaleh Kazeminejad, Zhenhailong Wang, Chris Callison-Burch, Carl Vondrick, Mohit Bansal, Dan Roth, Jiawei Han, Shih-Fu Chang, Martha Palmer, and Heng Ji. 2022b. Resin-11: Schema-guided event prediction for 11 newsworthy scenarios. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: System Demonstrations*.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3558–3565.

Octavian-Eugen Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic neural networks. *arXiv preprint arXiv:1805.09112*.

Hongyang Gao and Shuiwang Ji. 2019. Graph u-nets. In *international conference on machine learning*, pages 2083–2092. PMLR.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect

annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.

Tanya Goyal and Greg Durrett. 2019. Embedding time expressions for deep temporal ordering models. In *ACL*.

Mark Granroth and Stephen Clark. 2016a. What happens next? event prediction using a compositional neural network model. In *AAAI Conference on Artificial Intelligence*.

Mark Granroth and Stephen Clark. 2016b. What happens next? event prediction using a compositional neural network model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Jian Guan, Yansen Wang, and Minlie Huang. 2018. Story ending generation with incremental encoding and commonsense knowledge. In *AAAI Conference on Artificial Intelligence*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. 2016. Table filling multi-task recurrent neural network for joint entity and relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2537–2547, Osaka, Japan. The COLING 2016 Organizing Committee.

Rujun Han, I-Hung Hsu, Mu Yang, A. Galstyan, R. Weischedel, and Nanyun Peng. 2019a. Deep structured neural network for event temporal relation extraction. In *CoNLL*.

Rujun Han, Mengyue Liang, Bashar Alhafni, and Nanyun Peng. 2019b. Contextualized word embeddings enhanced event temporal relation extraction for story understanding. *ArXiv*, abs/1904.11942.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019c. Joint event and temporal relation extraction with shared representations and structured prediction. In *EMNLP/IJCNLP*.

Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. *ArXiv*, abs/2009.07373.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021a. Cuad: An expert-annotated nlp dataset for legal contract review. *ArXiv*, abs/2103.06268.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021b. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Rei Ikuta, Will Styler, Mariah Hamang, Timothy J. O'Gorman, and Martha Palmer. 2014. Challenges of adding causation to richer event descriptions. In *EVENTS@ACL*.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 1681–1691.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Bram Jans, Steven Bethard, Ivan Vulic, and Marie-Francine Moens. 2012a. Skip n-grams and ranking functions for predicting script events. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Bram Jans, Steven Bethard, Ivan Vulic, and Marie-Francine Moens. 2012b. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 336–344. ACL; East Stroudsburg, PA.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. 2020. Forecastqa: A question answering challenge for event forecasting with temporal text data. *arXiv preprint arXiv:2005.00792*.

Jaehyeong Jo, Jinheon Baek, Seul Lee, Dongki Kim, Minki Kang, and Sung Ju Hwang. 2021. Edge representation learning with hypergraphs. *Advances in Neural Information Processing Systems*, 34:7534–7546.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Transactions of the Association for Computational Linguistics*, 6:225–240.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Oleksandr Kolomiyets, Steven Bethard, and Marie Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97.

I-Ta Lee, Maria Leonor Pacheco, and Dan Goldwasser. 2020. Weakly-supervised modeling of contextualized event embedding for discourse relations. In *Findings*.

A. Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. *ArXiv*, abs/1808.09401.

A. Leeuwenberg and Marie-Francine Moens. 2019. A survey on temporal reasoning for temporal information extraction from text. *ArXiv*, abs/2005.06527.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare R. Voss. 2021a. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Conference on Empirical Methods in Natural Language Processing*.

Manling Li, Tengyu Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021b. Timeline summarization based on event graph compression via time-aware optimal transport. In *Conference on Empirical Methods in Natural Language Processing*.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare R. Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Conference on Empirical Methods in Natural Language Processing*.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. *arXiv preprint arXiv:1805.05081*.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining gcn and bert. *arXiv preprint arXiv:2105.05727*.

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

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

Zhengyuan Liu, Ke Shi, and Nancy F Chen. 2021. Coreference-aware dialogue summarization. *arXiv preprint arXiv:2106.08556*.

Shangwen Lv, Wanhui Qian, Longtao Huang, Jizhong Han, and Songlin Hu. 2019. Sam-net: Integrating event-level and chain-level attentions to predict what happens next. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6802–6809.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Puneet Mathur, Meghna Ayyar, Sahil Chopra, Simra Shahid, Laiba Mehnaz, and Rajiv Shah. 2018a. Identification of emergency blood donation request on twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 27–31.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. Timers: Document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533.

Puneet Mathur, Vlad Morariu, Verena Kaynig-Fittkau, Jiuxiang Gu, Franck Dernoncourt, Quan Hung Tran, Ani Nenkova, Dinesh Manocha, and Rajiv Jain. 2022. Doctime: A document-level temporal dependency graph parser. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 993–1009.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018b. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148.

Puneet Mathur, Ramit Sawhney, Shivang Chopra, Maitree Leekha, and Rajiv Ratn Shah. 2020. Utilizing temporal psycholinguistic cues for suicidal intent estimation. In *European Conference on Information Retrieval*, pages 265–271. Springer.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018c. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. Snap-batnet: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 conference of the North American Chapter of the association for computational linguistics: student research workshop*, pages 147–156.

Trisha Mittal, Puneet Mathur, Aniket Bera, and Dinesh Manocha. 2021. Affect2mm: Affective analysis of multimedia content using emotion causality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5661–5671.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and D. Roth. 2019a. An improved neural baseline for temporal relation extraction. In *EMNLP/IJCNLP*.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019b. An improved neural baseline for temporal relation extraction. *arXiv preprint arXiv:1909.00429*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020a. Torque: A reading comprehension dataset of temporal ordering questions. *arXiv preprint arXiv:2005.00242*.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020b. Torque: A reading comprehension dataset of temporal ordering questions. In *EMNLP*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *ACL*.

Yunseok Noh, Yong-Min Shin, Junmo Park, A.-Yeong Kim, Su Jeong Choi, Hyun-Je Song, Seongbae Park, and Seyoung Park. 2020. Wire: An automated report generation system using topical and temporal summarization. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Timothy J. O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation.

Karl Pichotta and Raymond J. Mooney. 2016. Using sentence-level lstm language models for script inference. *ArXiv*, abs/1604.02993.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003a. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The timebank corpus. *Proceedings of Corpus Linguistics*.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. Timedial: Temporal commonsense reasoning in dialog. *arXiv preprint arXiv:2106.04571*.

Nils Reimers, N. Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *ACL*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Hayley Ross, Jonathon Cai, and Bonan Min. 2020. Exploring contextualized neural language models for temporal dependency parsing. *arXiv preprint arXiv:2004.14577*.

Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Shah. 2021a. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6751–6762.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. 2020a. Voltage: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 167–175.

Ramit Sawhney, Puneet Mathur, Taru Jain, Akash Kumar Gautam, and Rajiv Shah. 2021b. Multitask learning for emotionally analyzing sexual abuse disclosures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4881–4892.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020b. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM international conference on multimedia*, pages 456–465.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Chao Shang, Peng Qi, Guangtao Wang, Jing Huang, Youzheng Wu, and Bowen Zhou. 2021. Open temporal relation extraction for question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Ke Shi, Zhengyuan Liu, and Nancy F Chen. 2020. An end-to-end document-level neural discourse parser exploiting multi-granularity representations. *arXiv preprint arXiv:2012.11169*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Jannik Strötgen and Michael Gertz. 2013a. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47:269–298.

Jannik Strötgen and Michael Gertz. 2013b. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Reid Swanson and Andrew S Gordon. 2008. Say anything: A massively collaborative open domain story writing companion. In *Joint International Conference on Interactive Digital Storytelling*, pages 32–40. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James F. Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In **SEMEVAL*.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. *ArXiv*, abs/1902.01390.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4070–4078, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Le Nguyen, Franck Dernoncourt, Bonan Min, and Thien Huu Nguyen. 2022. Document-level event argument extraction via optimal transport. In *FINDINGS*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *ArXiv*, abs/2211.07342.

Zhongqing Wang, Yue Zhang, and Ching Yun Chang. 2017a. Integrating order information and event relation for script event prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 57–67.

Zhongqing Wang, Yue Zhang, and Ching-Yun Chang. 2017b. Integrating order information and event relation for script event prediction. In *Conference on Empirical Methods in Natural Language Processing*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Ronald J. Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1:270–280.

Sean Wilner, Daniel Woolridge, and Madeleine Glick. 2021. Narrative embedding: Re-Contextualization through attention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1393–1405, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xianchao Wu, Ander Martínez, and Momo Klyen. 2018. Dialog generation using multi-turn reasoning neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2049–2059, New Orleans, Louisiana. Association for Computational Linguistics.

H. Xu, Wenlin Wang, W. Liu, and Lawrence Carin. 2018. Distilled wasserstein learning for word embedding and topic modeling. *ArXiv*, abs/1809.04705.

Jiarui Yao, Haoling Qiu, Bonan Min, and Nianwen Xue. 2020. Annotating temporal dependency graphs via crowdsourcing. In *EMNLP*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019a. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019b. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with chatgpt. *arXiv preprint arXiv:2304.05454*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020a. Big bird: Transformers for longer sequences. *ArXiv*, abs/2007.14062.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020b. Big bird: Transformers for longer sequences. In *NeurIPS*.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. *arXiv preprint arXiv:2106.03618*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuchen Zhang and Nianwen Xue. 2018a. Neural ranking models for temporal dependency structure parsing. *arXiv preprint arXiv:1809.00370*.

Yuchen Zhang and Nianwen Xue. 2018b. Structured interpretation of temporal relations. *arXiv preprint arXiv:1808.07599*.

Yuchen Zhang and Nianwen Xue. 2019. Acquiring structured temporal representation via crowdsourcing: A feasibility study. In **SEMEVAL*.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1630–1641.

Xinyu Zhao, Shih-ting Lin, and Greg Durrett. 2020. Effective distant supervision for temporal relation extraction. *arXiv preprint arXiv:2010.12755*.

Jianming Zheng, Fei Cai, Yanxiang Ling, and Honghui Chen. 2020a. Heterogeneous graph neural networks to predict what happen next. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 328–338.

Jianming Zheng, Fei Cai, Yanxiang Ling, and Honghui Chen. 2020b. Heterogeneous graph neural networks to predict what happen next. In *COLING*.

Bo Zhou, Yubo Chen, Kang Liu, Jun Zhao, Jiexin Xu, Xiaojian Jiang, and Qiuxia Li. 2022a. Generating temporally-ordered event sequences via event optimal transport. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1875–1884.

Yucheng Zhou, Xiubo Geng, Tao Shen, Guodong Long, and Daxin Jiang. 2022b. Eventbert: A pretrained model for event correlation reasoning. In *Proceedings of the ACM Web Conference 2022*, pages 850–859.

Fangqi Zhu, Jun Gao, Changlong Yu, Wei Wang, Chen Xu, Xin Mu, Min Yang, and Ruifeng Xu. 2023. A generative approach for script event prediction via contrastive fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14056–14064.