# Multi-Event Temporal Ordering by Event Order Ranking

Anonymous ACL submission


## Abstract

Extracting relationships and ranking the temporal order of document-level events is a challenging task in information extraction. Previous methods primarily considered the event pair as the basic unit for processing, ignoring holistic connection among all events and background information remaining in the rest text. To address these issues, we redefine the multi-event temporal ordering as Event Order Ranking(EORank) task, and introduce the Multi-Event Temporal Ranking(MEtR) model. EORank simultaneously focuses on all events within a document from a holistic perspective. We design order loss functions for MEtR, and our experimental results demonstrate their superior performance compared to other state-of-the-art models across EORank tasks of different settings.[1]


## 1 Introduction

Understanding the semantics and temporal relationships of events has been a long-standing fundamental task in natural language processing(Minsky, 1974;Schank and Abelson, 1975;Chen et al., 2021). Notably, many domains can benefit from the advancements in determining temporal relations of multiple events, such as the construction and reasoning of the knowledge graph(Li et al., 2020, Du et al., 2022), event prediction(Li et al., 2018), and making decisions(Sun et al., 2018).

Events in natural language, often represented by trigger words or the sentences containing them, construct a document as a story, wherein the underlying temporal relationships among them become notably intricate. The extraction of relations from these events scattered across the document is conventionally modeled as the Document-level Event-Event Relation Extraction(DERE) task(Yuan et al., 2023a;Tran Phu and Nguyen, 2021;Cohen and Bar, 2023) and subdivided into the Event Ordering task(McDowell et al., 2017; Chambers et al., 2014; Naik et al., 2019) to further focus on temporal relations. Moreover, events are intertwined for their temporal relations, which can be further ordered as chains(Zhang et al., 2021;Chambers and Jurafsky, 2008) by sequential ranking(Toro Isaza et al., 2023a).

Mostly, previous methods explore DERE and Event Ordering by identifying the relation in each event pair(Chambers and Jurafsky, 2008;Jans et al., 2012;Granroth-Wilding and Clark, 2016) with temporal information(Pichotta and Mooney, 2016b;Pichotta and Mooney, 2016a), decomposing the challenge of multiple events ordering into pairwise events temporal relation extraction subtasks (Ning et al., 2019;Zhang et al., 2022), which is considered a process of multi-class classification(Xiang and Wang, 2019) for each event pair.

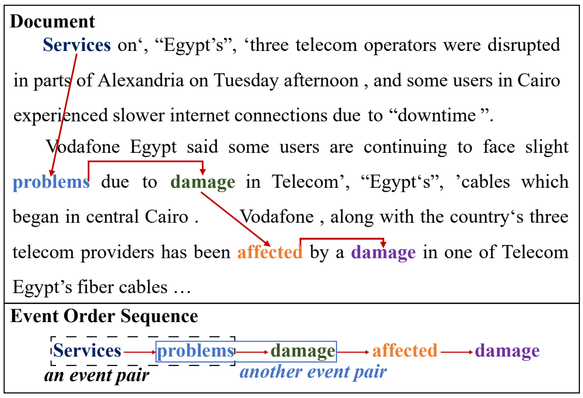



Figure 1: Example from EventStoryLine. Events are annotated by trigger words, and the order sequence is extracted from their relations. Some events and relations are omitted for clarity.

However, pairwise and multi-class classification methods also face defects(Examples in Appendix D). **(1)** Pairwise methods may predict a cyclical relationship among events without handling all events simultaneously, leading to a loop

[1]Code and data will be released after the review process.

that makes it impossible to determine the beginning and end of a story. **(2)** Classification methods may give the same classification result for different events that cause repetitive orders. Pairwise methods also need further processing to obtain the order sequence for events, such as constructing a relation adjacency matrix of each event pair separately and additional ranking algorithms to create sequential event chains(Toro Isaza et al., 2023a) from the matrix, which is computationally expensive and requires additional processes, reducing accuracy among multiple events.

While previous works focus on event pairs, real-world language often entails more than two events intertwined within a document, as shown in Figure 1. There is also information about events tangled with other events, narrated by the rest of the text, like an explanation of a specific concept, which can be considered background information(Hashimoto et al., 2014;Kruengkrai et al., 2017;Kadowaki et al., 2019). This fact of containing multiple events with other background information within a document is critical for DERE and Event Ordering. Thus, having a holistic perspective and obtaining the global event order sequence for multiple events is significant, rather than focusing on one local pair of events each time.

To handle events with background information from a holistic perspective and address defects of previous methods, we introduce the **E**vents **O**rder **R**anking(EORank) task and propose the **M**ulti-**E**vent Temporal **R**anking(MEtR) model. EORank approaches extracting temporal relationships among multiple events by holistically considering them as one cohesive story and requires the event order sequence for output. Diverging from prior tasks, we redefine the fundamental unit for event ordering as all events in the same text rather than event pairs. We aiming to simultaneously handle all information and circumvent the defects mentioned above by ranking the order of events at once.

The contributions of our paper are as follows:

- We introduce the **E**vents **O**rder **R**anking(EORank) task to achieve the multi-event temporal ordering ranking procedure with arranged data from datasets Event Storyline(Caselli and Vossen, 2017), ROCStory(Mostafazadeh et al., 2016) and StoryCommonsense(Rashkin et al., 2018). EORank requires ranking for event order sequence, enabling a holistic comprehension of multi-event relations.
- We propose **M**ulti-**E**vent temporal **R**anking(MEtR) model with two different loss functions, SOL and OCE, to address the EORank task, which handles all events simultaneously with a holistic perspective.
- Experimental results in EORank show that MEtR outperforms the baseline methods and demonstrates a remarkable effect in handling EORank tasks with more events.

## 2 Event Order Ranking

**Task Description** The task requiring ranking temporal order for events is referred to as Event Order Ranking (EORank). The primary objective of EORank is to predict the order $y_i$ for each event $e_i$, thereby forming the temporal order sequence $Y = \{y_i | i = 1, 2, \ldots, n\}$, *e.g.* considering events $e_1, e_2, e_3$, the predicted event order sequence $Y$ might be $\{y_1 = 0, y_2 = 2, y_3 = 1\}$ indicating the temporal orders as $e_1 \rightarrow e_3 \rightarrow e_2$.

In contrast with DERE, EORank does not constrained to rank an event chain from event pairs(Toro Isaza et al., 2023a), which allows for operating on multiple events simultaneously and ranking the event order sequence directly from the whole story. We enable cohesive comprehension of multi-event tasks for the first time through a holistic perspective in EORank, which does not necessarily require the basic process units to be event pairs.

**Dataset Arrangement** We modify the StoryCloze datasets and utilize the Event Storyline dataset to arrange data for EORank.

The original StoryCloze task entails selecting the correct ending from two candidate sentences for a background story consisting four sentences, each sentence represents an event and their temporal order aligns with their appearance. There are five events in a complete story from StoryCloze with the correct ending. We modify StoryCloze datasets(ROCStories and StoryCommonsense) for the EORank tasks by selecting and shuffling part of events in a story, and the objective is to rank them in order.

Furthermore, for more intricate situations, we utilize the dataset Event Storyline(ESL, Caselli and Vossen, 2017). ESL is a dataset that annotates events with trigger words, designed for temporal and causal relation detection among events. Unlike StoryCloze, where events are ordered coher-

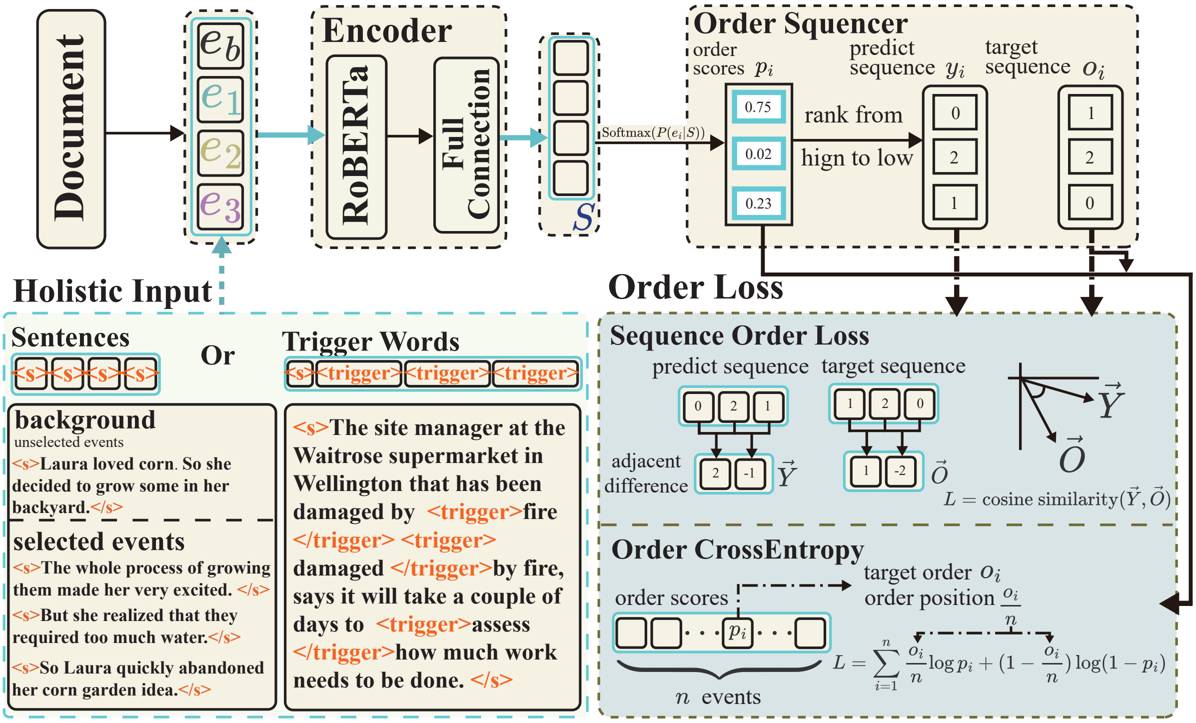



Figure 2: Overview of MEtR. Input data comprising information of the entire story are integrated in Holistic Input, incorporating both background(unselected events in StoryCloze and text other than trigger words in ESL) and events(annotated by special tokens,"**<s>**" and "**</s>**" for sentences or "**<trigger>**" and "**</trigger>**" for trigger words).

ently by their appearance in the story, events in ESL are not all directly narrated. Documents in ESL involve flashbacks and events intertwined with background information, offering more complex scenarios closer to real-world language. EORank task on ESL requires ranking temporal relation orders for selected events in each document.

## 3 MEtR

To handle all events simultaneously from a holistic perspective, we propose MEtR(Figure 2). MEtR inputs all events with background information simultaneously in Holistic Input(**HI**) structure and encodes them by RoBERTa(Liu et al., 2019) with a full connection layer. MEtR calculates order scores by Order Sequencer as output and ranks these scores into order sequences. Order scores of events reveal their temporal salience in the story. Order sequencer also reduces computation expense by predicting and ranking the probability order scores instead of treating all possible event orders as classification targets and sorting from a relation adjacency matrix of event pairs. To handle multi-event ordering with a holistic perspective, we devise the order sequencer with two different loss functions for MEtR.

### 3.1 Order Scores

MEtR considers a document a story $S$ consisting of interrelated events $e_i$ and background information $e_b$. Order sequencer takes $S$ with this interrelated information as the condition, defining the conditional probability $p_i$ of each event as order score:

$$p_i = \text{Softmax}(P(e_i|S)).$$

Like the coherence score from Granroth-Wilding and Clark, 2016 in event chains, order scores represent the confidence from MEtR for each event in a story. A higher order score signifies $e_i$ has a stronger correlation with $e_b$ and more salience in temporal relation within $S$ than other events with lower scores, indicating that $e_i$ should occur earlier(Figure 3). Order sequencer can determine the event order sequence for all events by ranking order scores from high to low.

### 3.2 Loss Functions

The objective of EORank is ranking events to obtain an ordered sequence $Y = y_1, y_2, \ldots, y_n$ where $y_i$ represents the predicted order of event $e_i$ while

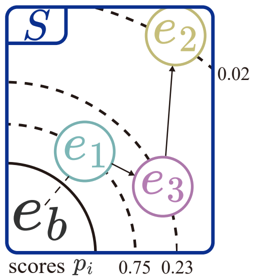



Figure 3: $p_i$ represents the correlation between event $e_i$ and background and its salience compared to other events in the story.

sequence $O = o_1, o_2, \ldots, o_n$ represents the true orders where $n$ is the number of events. MEtR outputs this sequence by ranking order scores $p_1, p_2, \ldots, p_n$ of each event from highest to lowest.

In order to obtain reasonable order scores, we devise two order loss functions: Sequence Order Loss(SOL) and Order Cross Entropy(OCE). We devise them from relative and absolute perspectives, respectively.

**Sequence Order Loss** For events, their ranked order sequence signifies the before-and-after relationship among them. Thus, SOL evaluates the order sequence by the relative direction and distance characteristics in sequence between adjacent events. The signed value of difference between orders of two events $o_i - o_j$ reflects the before-and-after direction and relative distance between them, *e.g.* if $o_i - o_j < 0$ then $e_i$ precedes $e_j$, and the absolute difference $|o_i - o_j|$ indicates the distance between $e_i$ and $e_j$. Remarkably, the order of initial and final events should give the greatest difference and direct from the initial to the final event.

These numerical and directional characteristics are consistent with the properties of vectors. Thus, we extract these relative characteristics of an order sequence as a vector $\vec{Y}$ by the adjacent signed value of order differences $o_{i+1} - o_i$. Based on these, we define SOL as:

$$
\begin{gathered}
L_{\text{SOL}} = \text{cosine similarity}(\vec{Y}, \vec{O}) \\
\vec{Y} = \{y'_i | y'_i = y_{i+1} - y_i,\ i = 1, 2, \cdots, n-1\} \\
\vec{O} = \{o'_i | o'_i = o_{i+1} - o_i,\ i = 1, 2, \cdots, n-1\},
\end{gathered}
$$

to depict the holistic characteristics of predicted order sequences and compare them with true orders.

**Order Cross Entropy** On the other hand, we introduce OCE to add absolute position information of each event into cross-entropy.

The order of an event tells its absolute position in the whole story, and explicitly associated with the total number of events. We describe this absolute position information of one event by its proportion of true order in sequence as $\frac{o_i}{n}$. This proportion reflects the absolute characteristics of one event in the sequence, and we replace the target(0 or 1) in binary cross-entropy with this proportion to assemble OCE. Thus, we design OCE as a variant of binary cross-entropy with absolute order proportion of each event:

$$
L_{\text{OCE}} = \sum_{i=1}^{n} \frac{o_i}{n} \log p_i + (1 - \frac{o_i}{n}) \log(1 - p_i),
$$

so that former events lead to higher order scores.

We utilize two loss functions separately for different perspectives with MEtR as MEtR$_{\text{SOL}}$ and MEtR$_{\text{OCE}}$. We also conduct comparative experiments with their summation.

## 4 Experiment

### 4.1 Datasets and Tasks

| Dataset | train | test | val |
|---|---|---|---|
| ROCStories | - | 1871 | 1871 |
| SC | 9885 | 2370 | 2483 |
| ESL | 1450 | 294 | - |

Table 1: Overview of dataset statistics. ROCStories and SC are datasets from the StoryCloze task.

We conduct experiments on StoryCloze to compare baselines and EORank tasks on StoryCloze and ESL datasets.

The original StoryCloze task entails a system selecting the correct ending for a multi-sentence story, where each sentence represents an event. As the original StoryCloze task is a binary classification task with only one number output, MEtR is not utilized in it. We conduct an experiment on StoryCloze datasets ROCStories(Mostafazadeh et al., 2016) and StoryCommonsense(SC, Rashkin et al., 2018) with baselines and other state-of-the-art methods to compare the capabilities of baselines in handling binary event tasks and lay a groundwork for comparing MEtR with baselines on more complex EORank tasks.

EORank tasks require ranking temporal order for events in a story. We select at least two events for ranking and, at most, five, for there are a total of five events in a story. We randomly shuffle these selected events, which need to be ranked, while the rest are considered as background. Specifically, there is no background with all five events selected.

ROCStories comprises a train dataset without incorrect options; thus, akin to prior works, we utilize its test dataset for training and the dev dataset for validation. StoryCommonsense is a modified version of ROCStories with additional annotations, providing a complete and abundant train dataset.

Further, for more complex scenarios, we also conduct EORank tasks by utilizing temporal relations of events in the Event Storyline(ESL, Caselli and Vossen, 2017) dataset for a comprehensive analysis.

Table 1 shows a summary of the statistics of these datasets.

### 4.2 Baselines

We design two baseline methods stand for previous pairwise and multi-class classification methods respectively. We choose the box embedding method as a pairwise method because of its superiority over previous pairwise methods. We also design a multi-class classification method to intuitively give orders for all events, demonstrating promising results in relation extraction between two events and ensuring specific effects in EORank.

**Box Model** Inspired by the box model(**B**ox **E**vent **R**elation **E**xtraction, BERE, Hwang et al., 2022a), we adapted BERE by employing box embeddings to EORank tasks, a typical pairwise method that takes RoBERTa as its encoder. BERE projects each event to a box representation which calculates the conditional probability $P(e_i \cap e_j | e_j)$ stands for $e_j \rightarrow e_i$ of each event pair to construct the relations matrix and rank the event order sequence.

The box embeddings make an event box contain anthor box related to it. This design initially intends to describe relations among multiple events, and BERE extracts pairwise relations by the relative position of the two boxes.

We train the model with the multi-event pairwise loss function:

$$-\sum \operatorname{sgn}(o_i - o_j)\Big[\log P(c_i \cap c_j | c_j) - \log P(c_i \cap c_j | c_i)\Big],$$

where sgn is sign function, $c_i$ is intersection of background $b$ with event $e_i$. More information about the BERE method is in Appendix C.

**Multi-class Classification Method** Inspired by Li et al., 2021, we also designed a **M**ulti-class **C**lassification **M**ethod(**MCM**) by assembling a Pair Input input layer into it and replacing BERT with RoBERTa as its encoder. MCM also utilizes RoBERTa as its encoder, like MEtR, to maintain the consistency of structure. Thus, both models have similar structures with different input and output layers.

The input structure of MCM is Pair Input(**PI**), which pairs events with the background respectively to integrate them. MCM is initially designed with background information, which can also integrate Holistic Input into it, in case the background may separate from events in the text or even not exist.

The output of MCM approaches the EORank task intuitively as multi-class classification, calculating the probability $P_i = P(y_i = t | e_i; S)$ for each event where $t$ represents each order number. While this structure ensures the efficiency of MCM, it may result in the defect of repetitive order in the output caused by separate classification of each event.

**GPT Prompt** Yuan et al., 2023b employ zero-shot prompt(Liu et al., 2023) on ChatGPT for EORank. Additionally, aiming to compare MEtR with a state-of-the-art LLM, we also employ GPT-3.5 by the prompt method in EORank tasks.

However, the effectiveness of prompt methods can vary based on the specific template design. While we select our prompt template after comparing results from various designs, it is crucial to acknowledge that the effectiveness of prompts can fluctuate with template variations. Further details regarding the prompt design can be found in Appendix B.

| Model | accuracy | F1 |
|---|---|---|
| TransBERT(Li et al., 2021) | 91.8% | - |
| GPT-3(Brown et al., 2020) | 83.2% (zero-shot)<br>87.7% (few-shot) | - |
| GraphBERT(Du et al., 2022) | 89.8% | - |
| BLOOMZ(Muennighoff et al., 2023) | 96.26% | - |
| BERE | 59.85% | 0.545 |
| MCM | 97.93% | 0.661 |

Table 2: Model performance on original StoryCloze dataset ROCStories. MEtR is a multi-event model and not in this task, while StoryCloze is a binary classification task.

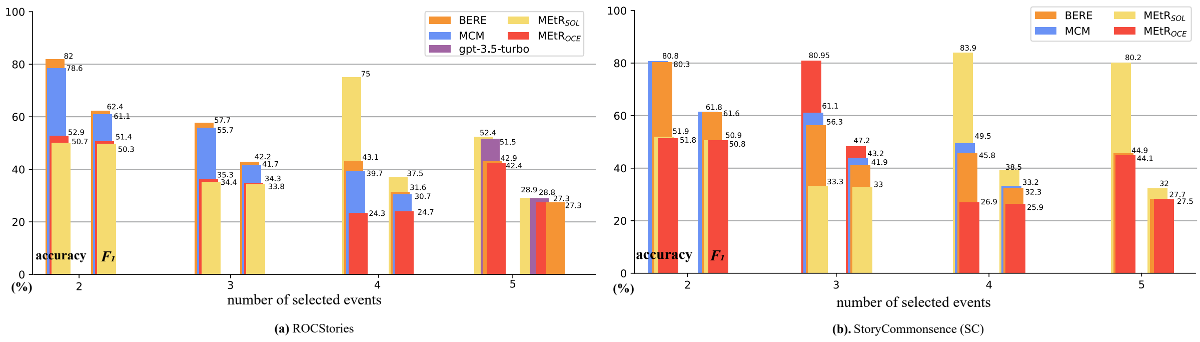



Figure 4: EORank task on StoryCloze datasets. MCM is not utilized in this task because there is no background information when all five events are selected.

### 4.3 Results and Analysis

We report the average accuracy(micro-$F_1$) to follow evaluation settings in previous works and additionally report macro-$F_1$ score to characterize the quality of event order ranking. More details about training settings can be found in Appendix A.

**Original StoryCloze** We evaluate the capabilities of baseline methods in handling binary event tasks on StoryCloze. Results presented in Table 2 showcase the performance of baselines and various other models. Since baseline MCM outperforms other methods in this task, migrating it to the newly proposed EORank task is promising.

TransBERT(Li et al., 2021) is structured similarly to MCM and utilizes BERT as the PLM. GraphBERT(Du et al., 2022), a method merging PLM and knowledge graphs, employs additional graph information to enhance PLM performance. BLOOMZ(Muennighoff et al., 2023) and GPT-3(Brown et al., 2020) are LLMs that have similar parameter sizes to each other. BLOOMZ undergoes multi-task prompted finetuning, while GPT-3 utilizes In-Context Learning(Brown et al., 2020;Dong et al., 2023).

Results on StoryCloze highlight the outstanding performance of MCM as a typical multi-class classification method, achieving a remarkable accuracy of 97.93% on ROCStories, which surpasses previous methods, including LLMs. Compared with TransBERT and GraphBERT, MCM utilizes superior RoBERTa as the encoder. Compared with LLMs BLOOMZ and GPT-3, MCM focuses on only one dataset with fewer parameters that can be fully finetuned. BERE is designed to handle multi-event tasks using the pairwise method, which may be less effective on binary StoryCloze but guarantees effective results on EORank tasks.

These results on binary event tasks lay the groundwork for comparing MEtR with baselines on more complex multi-event EORank tasks.

**EORank: StoryCloze** These tasks are based on StoryCloze datasets, with the number of events varying from 2 to 5, introducing increasingly complex scenarios among multiple events. The results of these tasks are in Figure 4.

Experiment results suggest that both baselines MCM and BERE have reliable capabilities in resolving EORank tasks, particularly in simpler scenarios with 2 and 3 events. Both MEtR$_{\text{SOL}}$ and MEtR$_{\text{OCE}}$ are weak with fewer events, especially with two events(31.3% accuracy gap at most, 2&3 events, Figure 4.a), for binary events task is close to classification task which baselines are excelling in, validated in original StoryCloze.

Meanwhile, MEtR$_{\text{OCE}}$ shows its superiority of handling fewer events with abundant training data on dataset ROCStories than SC, obtaining a 24.65% accuracy improvement to BERE(3 events, Figure 4.b). In more intricate scenarios, MEtR$_{\text{SOL}}$ shows a stable superiority with more events, obtains a 38.1% accuracy improvement at most with 6.2% $F_1$ improvement compared to BERE(4 events, Figure 4.b).

The difference between SOL and OCE in effectiveness comes from their different emphases. SOL focuses on the holistic order sequence, excelling with more events, while OCE emphasizes the absolute position of each event.

The prompt method employing gpt-3.5-turbo in the most intricate scenario with five events also suggests the remarkable effectiveness of zero-shot methods on LLMs in resolving EORank tasks, which falls behind MEtR$_{\text{SOL}}$ 0.9% at accuracy(5 events, Figure 4.a).

Among these EORank: StoryCloze tasks, MEtR shows superior results with abundant data, especially $\text{MEtR}_{\text{SOL}}$ is superior to $\text{MEtR}_{\text{OCE}}$ and baselines in more intricate scenarios.

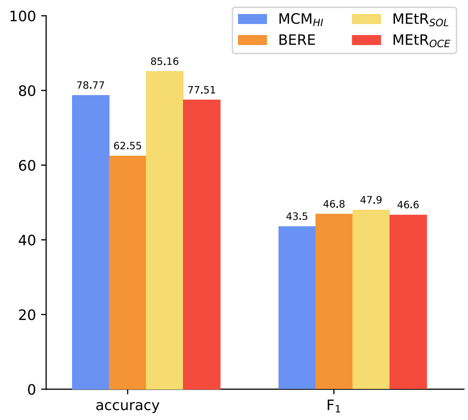



Figure 5: Model performance on Event Storyline(ESL). We employ Holistic Input for MCM because events are combined with background information.

**EORank: ESL** We also conduct EORank experiments on ESL for more intricate situations. For lack of data with more events in sequential temporal relations(Appendix A) and to maintain consistency with experiments of the original BERE method, we select three events in each ESL document for EORank.

Results in Figure 5 showcase that both $\text{MEtR}_{\text{SOL}}$ and $\text{MEtR}_{\text{OCE}}$ have promising capability in handling intricate conditions with annotation format utilizing trigger words. It is noticeable that BERE, as a pairwise method, yields a 62.55% accuracy, which maintains the same level as its original experimental(Hwang et al., 2022b), being inferior compared with MEtR and MCM.

In contrast, $\text{MEtR}_{\text{SOL}}$ significantly pushes the accuracy to 85.16%(22.61% accuracy and 1.1% $\text{F}_1$ improvement to BERE), which excels in trigger words annotated dataset comparing to results of EORank: StoryCloze. This improved effectiveness of models is caused by the better integration of events and text rather than separated sentences.

**Summation of Loss Functions** To further analyze the impact of loss functions, we conduct experiments with the summation of both loss functions $L = L_{\text{SOL}} + L_{\text{OCE}}$. $\text{MEtR}_{\text{SOL, OCE}}$ refers to the model trained with two loss function combined, and the results are in Table 3.

According to the volumes of different StoryCloze datasets, it is notable that dataset SC contains a larger quantity of data compared to ROCStory. Consequently, both loss functions demonstrate greater performances with increased data volume on SC.

Compared to each other, efficacy of loss functions varies with the number of events. In scenarios with fewer events like 3, OCE showcases specialized expertise as 14.2% $\text{F}_1$ improvement to SOL. Conversely, with five events, the 80.15% accuracy with 0.32 $\text{F}_1$ shows that $\text{MEtR}_{\text{SOL}}$ is excelling in EORank with more events.

Moreover, the summation of SOL and OCE encounters poor combinations as effectiveness decreases among EORank except for the task on StoryCommonsense with two events, which is closer to binary classification like the original StoryCloze. This poor combination stems from the different emphasis of two loss functions. The shortage of SOL with fewer events also highly affects the summation, giving a gap of 14.1% $\text{F}_1$ between $\text{MEtR}_{\text{SOL,OCE}}$ and $\text{MEtR}_{\text{OCE}}$(3 events).

It can be concluded that in EORank tasks, it is appropriate to utilize loss functions and methods separately in the scenarios in which they excel.

**Time Cost** Pairwise methods like BERE hanldle multi-event relations by constructing the matrix consists the probability of each event pair and ranking the final order sequence from the matrix, while MEtR is designed to rank output without the matrix, which reduces time cost.

The total time cost among EORank tasks of MEtR is 20.1% less than BERE(33.43% less on StoryCloze, 7.8% less on ESL).

**Error Analysis** MCM outputs orders by classification of each event separately and may output repetitive orders. BERE takes event pairs as process units and may even cause loops.

On average, in EORank tasks on StoryCommonsense, MCM outputs repetitive orders in 19.98% of cases, while BERE outputs loops in 0.43%(examples in Appendix D).

Meanwhile, MEtR interprets output by treating orders as a sequence naturally, ensuring no repetitive and loop output by the design of its order sequencer.

## 5 Conclusion

We address the challenge of multi-event temporal ordering from a cohesive perspective and circumvent defects caused by previous pairwise and multiclass classification methods. To reach these targets, we propose the EORank task to rank the temporal

| Model | Task on ROCStories | | | | | | | | Task on ESL | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 2 | | 3 | | 4 | | 5 | | | |
| | accuracy | F1 | accuracy | F1 | accuracy | F1 | accuracy | F1 | accuracy | F1 |
| MEtR$_{\text{SOL, OCE}}$ | 51.61% | 0.508 | 32.6% | 0.33 | 61.85% | 0.356 | **59.9%** | **0.3** | 84.36% | 0.478 |
| MEtR$_{\text{SOL}}$ | 50.70% | 0.503 | 34.36% | 0.338 | **75.03%** | **0.375** | 52.38% | 0.289 | **85.16%** | **0.479** |
| MEtR$_{\text{OCE}}$ | **52.91%** | **0.514** | **35.34%** | **0.343** | 24.30% | 0.247 | 42.36% | 0.273 | 77.51% | 0.466 |
| **Model** | **Task on StoryCommonsense** | | | | | | | | | |
| | 2 | | 3 | | 4 | | 5 | | | |
| | accuracy | F1 | accuracy | F1 | accuracy | F1 | accuracy | F1 | | |
| MEtR$_{\text{SOL, OCE}}$ | **92.13%** | **0.648** | 32.91% | 0.331 | 75.94% | 0.376 | 60.23% | 0.3 | | |
| MEtR$_{\text{SOL}}$ | 51.85% | 0.509 | 33.30% | 0.330 | **83.88%** | **0.385** | **80.15%** | **0.320** | | |
| MEtR$_{\text{OCE}}$ | 51.77% | 0.508 | **80.95%** | **0.472** | 26.88% | 0.259 | 44.05% | 0.275 | | |

Table 3: Comparison of loss functions.

order for events and the MEtR model to handle all events simultaneously with less computation by its holistic input structure and order sequencer. Experimental results demonstrate the effectiveness of the devised loss functions, SOL and OCE, showcasing their specialization in scenarios with various experimental settings. In contrast to other state-of-the-art methods, even LLMs, MEtR outperforms them in intricate multi-event EORank tasks, demonstrating superior performance.

## Limitations

A key limitation in our work is not addressing simultaneous temporal relations and no-relation within MEtR for the reasons below:

- We maintain consistency with previous works for interpreting the holistic narrative plot(Toro Isaza et al., 2023b).
- Datasets ROCStories and SC lack these specific types of relation. To maintain consistency, we exclude these data from the ESL dataset.
- We perceive the extraction of simultaneous temporal relations and no-relation between two events as subtasks, ideally performed after obtaining the temporal order sequence by measuring the adjacent events in that order. Notably, various established methods, such as box embeddings, are proficient in handling these relations, suggesting a potential avenue for future work.

## References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. 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, Columbus, Ohio. Association for Computational Linguistics.

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021. Event-centric natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 6–14, Online. Association for Computational Linguistics.

Omer Cohen and Kfir Bar. 2023. Temporal relation classification using Boolean question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1843–1852, Toronto, Canada. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Li Du, Xiao Ding, Yue Zhang, Ting Liu, and Bing Qin. 2022. A graph enhanced BERT model for event prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2628–2638, Dublin, Ireland. Association for Computational Linguistics.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. 2022a. Event-event relation extraction using probabilistic box embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 235–244, Dublin, Ireland. Association for Computational Linguistics.

EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. 2022b. Event-event relation extraction using probabilistic box embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, page 239, Dublin, Ireland. Association for Computational Linguistics.

Bram Jans, Steven Bethard, Ivan Vulić, and Marie Francine Moens. 2012. 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*, pages 336–344, Avignon, France. Association for Computational Linguistics.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. 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 5816–5822, Hong Kong, China. Association for Computational Linguistics.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. *CoRR*, abs/1805.05081.

Zhongyang Li, Xiao Ding, and Ting Liu. 2021. Transbert: A three-stage pre-training technology for story-ending prediction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(1).

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

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

Bill McDowell, Nathanael Chambers, Alexander Ororbia II, and David Reitter. 2017. Event ordering with a generalized model for sieve prediction ranking. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 843–853, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Marvin Minsky. 1974. A framework for representing knowledge.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. 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 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Karl Pichotta and Raymond Mooney. 2016a. Statistical script learning with recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 11–16, Austin, TX. Association for Computational Linguistics.

Karl Pichotta and Raymond J. Mooney. 2016b. Using sentence-level LSTM language models for script inference. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–289, Berlin, Germany. Association for Computational Linguistics.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299, Melbourne, Australia. Association for Computational Linguistics.

Roger C Schank and Robert P Abelson. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, pages 151–157.

Yawei Sun, Gong Cheng, and Yuzhong Qu. 2018. Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 806–817, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023a. Are fairy tales fair? analyzing gender bias in temporal narrative event chains of children's fairy tales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6509–6531, Toronto, Canada. Association for Computational Linguistics.

Paulina Toro Isaza, Guangxuan Xu, Toye Oloko, Yufang Hou, Nanyun Peng, and Dakuo Wang. 2023b. Are fairy tales fair? analyzing gender bias in temporal narrative event chains of children's fairy tales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 6520, Toronto, Canada. Association for Computational Linguistics.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490, Online. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Changsen Yuan, Heyan Huang, Yixin Cao, and Yonggang Wen. 2023a. 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)*, pages 16222–16234, Toronto, Canada. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023b. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390, Seattle, United States. Association for Computational Linguistics.

Xiyang Zhang, Muhao Chen, and Jonathan May. 2021. Salience-aware event chain modeling for narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1418–1428, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

# A MEtR Training

**Hyperparameters** We employ AdamW as the optimizer and utilized a cosine scheduler with hard restarts for each cycle during the training of MEtR. Notably, we observe that the performance of models exhibits instability with higher learning rates, particularly with an increased number of events.

Given this instability, we deliberately select lower learning rates to ensure more stable training results. The recommended learning rates for MEtR in EORank:ESL is $5 \times 10^{-6}$, $5e-6$, and the settings in EORank:StoryCloze shown in Table 4. Details about the code training logs can be found in our github page.[2]

It is essential to note that reducing the learning rate might necessitate increasing in the number of training epochs to ensure models are fully trained. Technically, the size of MEtR is similar to previous PLM models with RoBERTa-large, whose training duration on each task typically falls within one day on 1*V100 under the provided settings.

| Task | Model | Dataset | |
|---|---|---|---|
| | | learning rate | |
| | | ROCStories | SC |
| 2 | $\text{MEtR}_{\text{SOL}}$ | 7e-7 | 6e-6 |
| | $\text{MEtR}_{\text{OCE}}$ | 6e-6 | 6e-6 |
| | $\text{MEtR}_{\text{PI,SOL}}$ | 5e-6 | 6e-6 |
| | $\text{MEtR}_{\text{PI,OCE}}$ | 5e-6 | 6e-6 |
| 3 | $\text{MEtR}_{\text{SOL}}$ | 7e-7 | 6e-6 |
| | $\text{MEtR}_{\text{OCE}}$ | 6e-6 | 7e-7 |
| | $\text{MEtR}_{\text{PI,SOL}}$ | 4e-6 | 5e-6 |
| | $\text{MEtR}_{\text{PI,OCE}}$ | 6e-6 | 5e-6 |
| 4 | $\text{MEtR}_{\text{SOL}}$ | 4e-6 | 6e-6 |
| | $\text{MEtR}_{\text{OCE}}$ | 4e-6 | 6e-6 |
| | $\text{MEtR}_{\text{PI,SOL}}$ | 5e-6 | 5e-6 |
| | $\text{MEtR}_{\text{PI,OCE}}$ | 6e-6 | 5e-6 |
| 5 | $\text{MEtR}_{\text{SOL}}$ | 5e-6 | 8e-7 |
| | $\text{MEtR}_{\text{OCE}}$ | 5e-6 | 7e-7 |

Table 4: Learning Rates utilized in EORank:StoryCloze.

**Numbers of events in ESL** In ESL, the data volume with 4 and 5 events is 934 and 479(total of train, test, and val) which significantly less than other task settings(Table 5). Thus, we utilize ESL only with 3 events in experiments.

## B Prompt Design

We choose a prompt template for gpt-3.5-turbo, designed as a step-by-step procedure. More details of the code and prompt in github page.[2]

[2]Will be released after the review process.

| Dataset | Volume |
|---|---|
| ROCStories | 3742 |
| StoryCommonsense | 14738 |
| ESL (3 events) | 1744 |
| ESL (4 events) | 934 |
| ESL (5 events) | 479 |

Table 5: Total volume of datasets.

## C Box Model

We train a box model with a pairwise loss function following the principles and techniques detailed in the BERE method.

For comprehensive details regarding the code we utilize, including implementation specifics and references, more information can be found on our GitHub page and the official page of box embeddings (`https://www.iesl.cs.umass.edu/box-embeddings/main/index.html`) for a deeper understanding of the methodology.

## D Case Study

We display the input text and output of some case examples from baselines in Table 6 for a better understanding of EORank tasks.

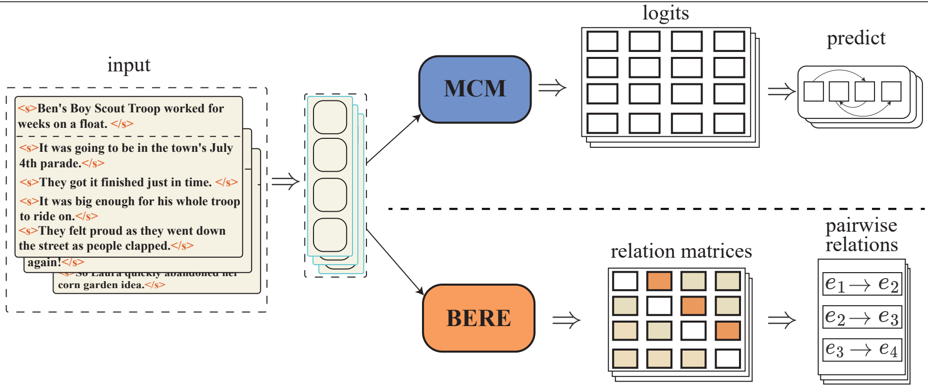
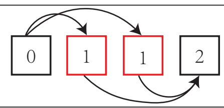
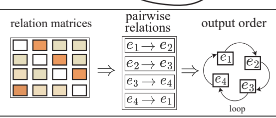
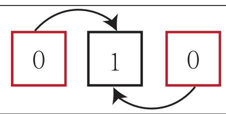
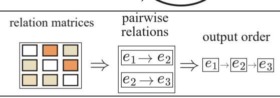
<table>
<tr><td colspan="4">
</td></tr>
<tr><th>Input(StoryCloze style)</th><th colspan="2">Output</th><th>Description</th></tr>
<tr><td rowspan="3">background:<br>Ben's Boy Scout Troop worked for weeks on a float.<br><br>events:<br>$e_1$: It was going to be in the town's July 4th parade.<br><br>$e_2$: They got it just in time.<br><br>$e_3$: It was big enough for his whole troop to ride on.<br><br>$e_4$: They felt proud as they went down the street as people clapped.</td><td colspan="2">$\boxed{0} \to \boxed{1} \to \boxed{2} \to \boxed{3}$<br><br>$e_1$: It was going to be...<br>⇓<br>$e_2$: They got it finished...<br>⇓<br>$e_3$: It was big enough...<br>⇓<br>$e_4$: They felt proud as...</td><td>Correct order</td></tr>
<tr><td>MCM</td><td>
</td><td>Repetitive orders cause branches.</td></tr>
<tr><td>BERE</td><td>
</td><td>Error pairwise relations cause loop.</td></tr>
<tr><th>Input(ESL style)</th><th colspan="2">Output</th><th>Description</th></tr>
<tr><td rowspan="3">A jury has handed down a guilty verdict for one of the two men accused of murdering ($e_1$) accused ($e_2$) of murdering a Brownsville mother of 12. Andrew Lopez was found guilty of second-degree murder and two counts of first-degree assault in the death of Zurana Horton. Jury deliberations ($e_3$) are still underway for the second suspect, Jonathan Carrasquillo. Horton was shot and killed in October 2011 when gunfire erupted outside...</td><td colspan="2">$\boxed{0} \to \boxed{1} \to \boxed{2}$<br>A jury has handed down a guilty verdict for one of the two men accused of murdering accused of murdering a Brownsville mother of 12. Andrew Lopez was found guilty of second-degree murder and two counts of first-degree assault in the death of Zurana Horton. Jury deliberations are still underway for the second suspect, Jonathan Carrasquillo. Horton was shot and killed in October 2011 when gunfire erupted outside...</td><td>Correct order.</td></tr>
<tr><td>MCM</td><td>
</td><td>Repetitive orders cause opposite branches.</td></tr>
<tr><td>BERE</td><td>
</td><td>Correct output.</td></tr>
</table>

Table 6: Case examples.