Extracting Military Event Temporal Relations via Relative Event Time Prediction and Virtual Adversarial Training

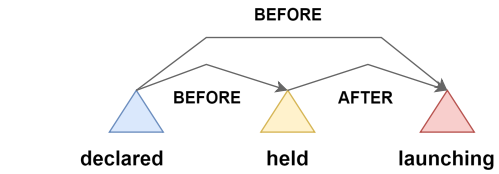
Jie Gong[1*], **Qiwang Hu**[2]
[1]School of Computer Science, South China Normal University
[2]School of Computer Engineering and Science, Shanghai University
gongjie@m.scnu.edu.cn, hqw17@shu.edu.cn

Abstract

Extracting temporal relationships between events in the text is crucial for understanding how events unfold over time, especially in the information-dense and precision-demanding military field. Existing models for extracting event temporal relations typically compare the relative times of events directly, neglecting the contextual information between event pairs. This can lead to difficulties in handling uncertain temporal boundaries expressed in text. In this paper, we propose an event temporal relationship extraction model for the military field, based on relative event time prediction and virtual adversarial training, MFRV. The relative event time prediction as an auxiliary task enhances the model's ability to capture and infer temporal relationships. Virtual adversarial training increases the model's generalization by generating adversarial samples. Additionally, we adopt the MoCo (Multi-objective gradient correction) method to balance the losses from relative event time prediction and virtual adversarial training, effectively resolving the gradient bias issue in multi-objective optimization. Furthermore, we have constructed a new dataset, TRMF, specifically for event temporal relationship extraction in the military field. Experiments conducted on TRMF, as well as widely used public datasets MATRES and TCR, demonstrate the effectiveness of MFRV. [1]

1 Introduction

In recent years, the increasing frequency of military events has garnered significant global attention, presenting considerable challenges to society. Unlike general news, military events are characterized by more complex dynamics across temporal and spatial dimensions, involving multiple stakeholders and interweaving of political, economic, and strategic factors. Accurately extracting temporal relationships from these events is critical for informed decision-making in policy, strategy formulation, and crisis management. However, errors in the timeline can disrupt the causal chain of events, potentially misleading analysts' understanding of the relationships between them. For example, an incorrect ordering of events could misinterpret the true cause-and-effect dynamics at play, thus distorting the strategic analysis.



One man held a sign that read, Russia will save the Donbas from war, referring to the area in southeast Ukraine that Russia declared independent this week before launching its invasion.

Figure 1: Example Sentences for Event Temporal Relation Extraction.

Such inaccuracies can have serious consequences for downstream tasks, including military action prioritization and intelligence analysis, where decisions might be made based on flawed temporal information, leading to incorrect judgments about the sequence and significance of events. In addition, military events often span extended timeframes, involve numerous actors, and may include classified or misleading information, rendering general temporal extraction techniques inadequate. To address this issue, we propose MFRV, a specialized model for extracting temporal relationships from military events.

Event Temporal Relationship Extraction is a natural language processing task that involves identifying and classifying the temporal order between pairs of events mentioned within a text. The goal is to determine the chronological sequence or temporal relationship between these events based on

[1]Experimental source code is available at https://github.com/Author0001/MFRV.

*Corresponding author.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 3305–3317

contextual and linguistic cues present in the text. This task is crucial for understanding the temporal dynamics in narrative and informative texts, enabling applications in timeline generation, question answering, and information retrieval (Choubey and Huang, 2017; Han et al., 2019b; Ning et al., 2020). For example, in Figure 1, there are three events with a clear temporal logic: "declared" occurs before "held," "held" occurs after "launching" and "declared" occurs before "launching". These events form a complete temporal relationship chain.

Neural network-based methods have achieved promising results in the field of temporal relationship extraction (Meng et al., 2017; Meng and Rumshisky, 2018; Ning et al., 2019; Han et al., 2019a; Cheng et al., 2020; Wang et al., 2020; Ballesteros et al., 2020; Tan et al., 2021; Zhang et al., 2022; Hwang et al., 2022; Tan et al., 2023). These methods mainly treat this task as a classification problem. Some methods focus on global structures, which encompass the overall temporal framework and the interconnections among multiple events within a text. These methods aim to capture the broader temporal context and dependencies that extend beyond individual event pairs (Bramsen et al., 2006; Yoshikawa et al., 2009; Zhou et al., 2022; Tan et al., 2024). Other works focus on event time information, such as the study by Leeuwenberg and Moens (2018). They proposed a method that infers temporal relationships by directly comparing the relative timestamps of events, thereby predicting relative timelines. Inspired by their work, Wen and Ji (2021) linked relative event time prediction with temporal relationship extraction by incorporating relative event time as additional features into the classifier's training. Although the model achieved promising performance, it lacked external knowledge assistance and demonstrated poor generalization to unseen data.

In this paper, we follow the ideas of Wen and Ji (2021), Leeuwenberg and Moens (2018) by predicting relative event times and incorporating them as additional features into the classifier's training. In predicting relative event times, we employed a dynamic adjustment mechanism to control the flow of information. This mechanism helps the model extract, filter, and utilize critical information closely related to temporal relationships. Additionally, we integrate external temporal commonsense knowledge (Ning et al., 2018b) during classifier training to improve the model's understanding of event temporal relationships. We also employ the virtual adversarial training (VAT) method (Miyato et al., 2018), generating adversarial samples to enhance the model's generalization. Furthermore, to effectively integrate the losses from relative event time prediction and virtual adversarial training, we use the MoCo method (Fernando et al., 2023) to balance these losses, resulting in improved model performance. Our contributions can be summarized as follows:

- We propose MFRV, an event temporal relationship extraction model based on relative event time prediction and virtual adversarial training, tailored for the military field.
- We construct TRMF, a novel dataset specifically designed for event temporal relationship extraction in the military field, addressing the data scarcity issue in this field.
- We conduct experiments on the TRMF, MATRES, and TCR datasets. The experimental results demonstrate the effectiveness of MFRV.

2 Related Work

Similar to entity-level relationship extraction (Peng et al., 2017; Li et al., 2019; Zhao et al., 2023), the latest event temporal relationship extraction models are based on neural networks (Zhuang et al., 2023b; Knez and Žitnik, 2024; Yuan et al., 2024). Specifically, Tan et al. (2023) combined the principles of statistical inference with deep learning methods. They employed Bayesian approaches to model the uncertainty in relationship extraction, thereby providing more reliable predictions. Hwang et al. (2022) used Box Embeddings to represent each event as a probabilistic box, which captures and models the complex temporal and spatial relationships between different events. Zhang and Li (2023) proposed a contrastive optimization-based framework. They trained the model by minimizing the discrepancy between model outputs and logical constraints. This enabled the model to better understand and capture temporal relationships. The most relevant work is by Wen and Ji (2021), which jointly trained relative time prediction and temporal relationship extraction. They used a stack propagation framework to utilize relative event times as features. Although the model achieved promising performance, it lacked external knowledge assistance and demonstrated poor

generalization to unseen data. Our work focuses on accurately predicting relative event times and enhancing the model's generalization capability by incorporating virtual adversarial training.

3 Approach

The architecture of the MFRV, as shown in Figure 2, comprises two modules: the Word Embedding Module and the Relative Event Time Knowledge Concatenation Module. The Word Embedding Module calculates word embeddings through event mentions and trigger words. The Relative Event Time Knowledge Concatenation Module uses the obtained embeddings for relative time prediction. It then combines the relative time and external knowledge with the trigger word embeddings to form the complete event embeddings. Finally, the temporal relationships between event pairs are predicted through classification. During training, the MFRV employs a combination of virtual adversarial training and the MoCo method to correct biased noisy gradient directions.

3.1 Word Embedding Module

The input to the Word Embedding Module is a segment of text containing event mentions with trigger words. This module utilizes the pre-trained language model Roberta-base (Liu et al., 2019) to compute the word embeddings of the text.

$$H = (h_1, h_2, \ldots, h_i, \ldots, h_j, \ldots, h_k) \quad (1)$$

where h_k represents the word embedding at position k.

3.2 Relative Event Time Knowledge Concatenation Module

When extracting event temporal relationships, it is intuitive to map event occurrence times to specific timestamps and then determine their temporal relationships by comparing the exact chronological order of two events (Leeuwenberg and Moens, 2018). Based on this reasoning, to better utilize the contextual information of events, we employ relative event time prediction. We predict event times as real numbers between -1 and 1. Using the pretrained language model, we obtain the word embeddings of the event pairs. These embeddings are processed through two feedforward neural networks (FFN_1 and FFN_2), and the dynamic adjustment mechanism is used to control the flow of information. Ultimately, we obtain the relative time of the event pairs.

$$\begin{aligned} t_i = \ & \tanh(\mathrm{FFN}_2(\tanh(\mathrm{FFN}_1(h_i)))) \\ & \cdot \sigma(\mathrm{FFN}_2(\tanh(\mathrm{FFN}_1(h_i)))) \end{aligned} \quad (2)$$

$$\begin{aligned} t_j = \ & \tanh(\mathrm{FFN}_2(\tanh(\mathrm{FFN}_1(h_j)))) \\ & \cdot \sigma(\mathrm{FFN}_2(\tanh(\mathrm{FFN}_1(h_j)))) \end{aligned} \quad (3)$$

where σ denotes the sigmoid. For events e_i and e_j, if their temporal relationship is "BEFORE", then their relative times t_i and t_j should satisfy $t_i < t_j$. If their temporal relationship is "AFTER", then their relative times t_i and t_j should satisfy $t_i > t_j$. If their temporal relationship is "EQUAL", then their relative times t_i and t_j should be as close as possible. Therefore, we use the following loss function to optimize the relative event time prediction.

$$\begin{aligned} L_t = & \ \partial[R_{(e_i,e_j)} = \text{BEFORE}] \max(0, 1 - (t_j - t_i)) \\ & + \partial[R_{(e_i,e_j)} = \text{AFTER}] \max(0, 1 - (t_i - t_j)) \\ & + \partial[R_{(e_i,e_j)} = \text{EQUAL}] \left| t_i - t_j \right| \end{aligned} \quad (4)$$

where $R_{(e_i,e_j)}$ represents the true temporal relationship between events e_i and e_j, and $\partial[\cdot]$ is an indicator function. The output value is 1 if the condition is true, and 0 if the condition is false.

After obtaining the relative time of the event pair, we further incorporate this feature into the event temporal relation extraction.

$$s_i = h_i \oplus t_i; s_j = h_j \oplus t_j \quad (5)$$

The MFRV enhances its ability to understand complex event relationships by incorporating the TEMPROB knowledge base (Ning et al., 2018b). TEMPROB is a comprehensive database of temporal commonsense knowledge that meticulously records the temporal relationships between numerous event pairs.

For example, the likelihood of an ask event occurring before the $help$ event is 86%, while the likelihood of it occurring after is only 9%. Similarly, for die and $explode$ events, the likelihood of die happening after $explode$ is 83%, while the likelihood of it happening before is 14%. These statistics provide us with a deep understanding of temporal event relationships. We train temporal commonsense knowledge encoding on this temporal commonsense knowledge base. The resulting knowledge vectors v_i and v_j are ultimately incorporated into the event temporal relationship extraction process.

$$w_i = s_i \oplus v_i; w_j = s_j \oplus v_j \quad (6)$$

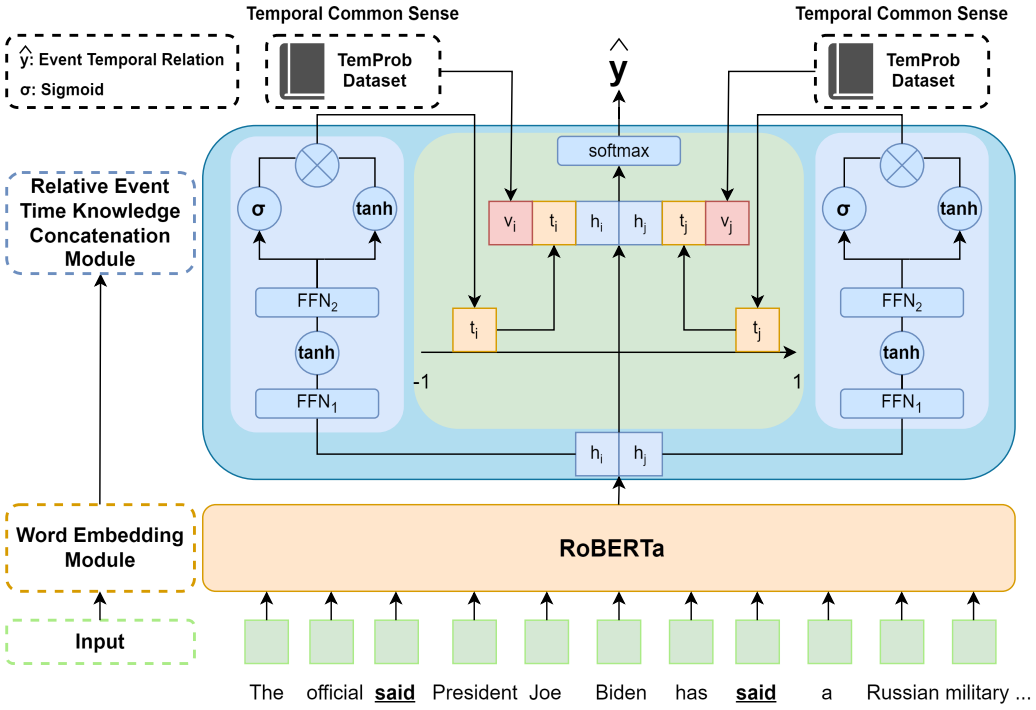


Figure 2: The Architecture of the MFRV.

Finally, the obtained word embeddings are concatenated to form the input $a_{(i,j)}$ for the softmax function. The probability distribution over the different classes is then computed to obtain the predicted temporal relationship P for the event pair e_i and e_j.

$$a_{(i,j)} = \text{FFN}_3\left(\tanh\left(\text{FFN}_4\left([w_i; w_j]\right)\right)\right) \quad (7)$$

$$P(y = R \mid x = (e_i, e_j)) = \text{softmax}(W a_{(i,j)} + B) \quad (8)$$

where W and B are the weight matrix and bias term parameters, respectively.

3.3 Virtual Adversarial Training

During the training phase, the MFRV employs the cross-entropy loss function as the optimization objective to accurately extract event temporal relationships.

$$L_r = -\log P(y = R \mid x = (e_i, e_j)) \quad (9)$$

Additionally, by incorporating the relative event time prediction task, we obtain the model's loss function, with β being the weight used to balance the two components.

$$L_a = L_r + \beta L_t \quad (10)$$

During the training process, we introduced virtual adversarial training combined with the MoCo method, as illustrated in Figure 3. We use the MoCo method to balance L_a and L_v, where L_v is obtained through virtual adversarial training. Virtual adversarial training (Miyato et al., 2018) is a regularization method that effectively enhances model robustness by promoting local smoothness of the model's posterior distribution. Specifically, we maximize the Kullback-Leibler (KL) divergence between the posterior distributions of the unperturbed and perturbed inputs. This approach emphasizes that the network's output should remain as consistent as possible under perturbations. We first compute the perturbation vector δ.

$$\delta = \arg\max_{\|\delta\| < \epsilon} \text{KL}\left(P(y = R \mid x = (e_i, e_j)) \,\|\, P(y = R \mid x = (e_i + \delta_i, e_j + \delta_j))\right) \quad (11)$$

where ϵ constrains the maximum perturbation magnitude, ensuring that the noisy input $x + \delta$ remains within an ϵ-radius around x. $\delta = \delta_i + \delta_j$ denotes the perturbation to the input. The smaller the KL divergence, the smoother the model's posterior distribution around x. The virtual adversarial training

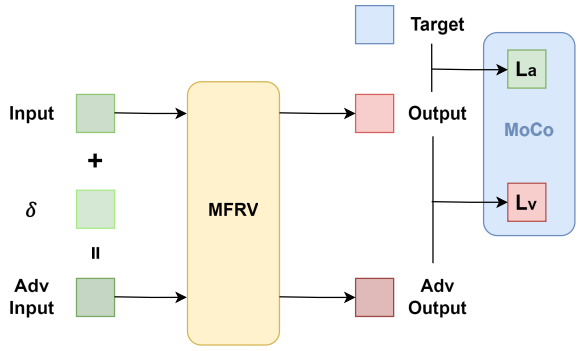


Figure 3: Training Process.

loss function is defined as follows.

$$L_v = \mathrm{KL}\left(P(y = R \mid x = (e_i, e_j)) \parallel P(y = R \mid x = (e_i + \delta_i, e_j + \delta_j))\right) \quad (12)$$

Next, we use the MoCo method to coordinate the optimization of the loss functions L_a and L_v. First, we define a gradient tracking variable $y_{(k,m)}$ for each loss function. Here, k denotes the training step or iteration number, specifically referring to the gradient update performed at step k. m represents different loss functions, serving as an index to distinguish between the different loss functions L_a and L_v. These variables are used to approximate the true gradients of each loss function and are updated at each training step. For each loss function L_m, we compute its gradient $h_{(k,m)}$, and then update the gradient tracking variable $y_{(k,m)}$ to reduce bias.

$$y_{(k+1,m)} = \Pi_{\Lambda_m}\left(y_{(k,m)} - \beta_k\left(h_{(k,m)} - y_{(k,m)}\right)\right) \quad (13)$$

where β_k represents the step size, and Π_{Λ_m} denotes the projection of the updated tracking variable onto the appropriate set Λ_m to ensure the reasonableness of the updates. Subsequently, λ is defined as the optimal scalar used to scale the gradient of each loss function and is updated via regularized stochastic projected gradient descent.

$$\lambda_{k+1} = \Pi_{\Lambda}\left(\lambda_k - \gamma_k\left(Y_k^T Y_k + \rho I\right)\lambda_k\right) \quad (14)$$

Updating the value of λ allows for the dynamic adjustment of the weights of the loss functions L_a and L_v during the optimization process. Here, γ_k represents the step size, I is the identity matrix, ρ is the regularization term, and Y_k is the approximation of the loss function's gradient. Finally, the model parameters x are updated.

$$x_{k+1} = \Pi_X\left(x_k - \alpha_k Y_k \lambda_k\right) \quad (15)$$

where α_k is the learning rate, which controls the step size of the parameter updates. Π_X denotes the projection onto the set X. These steps are repeated until a convergence criterion is met or a predetermined number of iterations is reached.

In each iteration, $y_{(k,m)}$ provides information about the current gradient estimates of each loss function, while λ ensures that these gradients are reasonably scaled and combined during the model parameter updates.

4 Experiments

This section provides a comprehensive introduction to the dataset we constructed, along with the datasets used in our experiments. It details the relevant experimental settings and presents the experimental results. Additionally, it includes an analysis of these results.

4.1 Constructing the TRMF Dataset

Considering the quality and authenticity of military filed texts, we used military news reports related to the Russia-Ukraine war from Voice of America (VOA)(https://www.voanews.com/) as our raw corpus. VOA is an American international broadcaster, the largest and oldest international broadcaster funded by the U.S. government. The VOA website features a Ukraine section. Leveraging VOA's military news data, we can easily access military news reports related to the Russia-Ukraine war. Similar to the MATRES dataset, we categorized the event temporal relations in TRMF into four types: BEFORE, AFTER, EQUAL, and VAGUE. Before data annotation, we performed the following preprocessing on the filtered raw corpus:

1) Sentence Segmentation: We used delimiters such as commas, question marks, semicolons, etc., to split the collected military news text into sentences.

2) Sentence Refinement: After segmentation, many new sentences are generated. We need to clean these sentences by removing empty, excessively long or short, and duplicate sentences. This is necessary to ensure the quality of the annotated corpus.

We used a large language model to first identify event trigger words and then annotate the event temporal relations. Finally, the annotations were manually reviewed. Appendix A shows the prompts we used in the extraction of event temporal relations.

A total of eight individuals (including two ex-

Label Type	TRMF			MATRES			TCR
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Test**
BEFORE	11694	2676	2570	5483	942	427	1780
AFTER	23055	4780	4659	3819	662	271	862
EQUAL	569	108	145	359	59	30	4
VAGUE	190	44	47	1227	189	109	0
TOTAL	35508	7608	7421	10888	1852	837	2646
TOTAL (SUM)		50537			13577		2646

Table 1: Annotation of Temporal Event Relationships in Three Datasets

perts) participated in this process. For each document, we required the participants to review it independently. If any issues arose, they could seek assistance from the experts. After each batch of documents was reviewed, the experts checked them. Documents that did not meet the standards were sent back for re-review. This process was repeated until the acceptance rate reached 90%. We used Fleiss' Kappa to calculate the Inter-annotator agreement scores, which yielded a score of 0.67.

Example from the dataset TRMF are shown below: *"Russia **denies** it is planning an **attack**, but says it could take unspecified military action if a list of demands is not met."*

In this sentence, "denies" is the trigger word for the denial event, and "attack" is the trigger word for the attack event. Based on the context, we can infer that the denial event occurs before the attack event.

Finally, we obtained 5800 documents. A comparison of the data with the MATRES and TCR datasets is shown in Table 1. The training set includes 35,508 instances, the development set contains 7,608 instances, and the test set consists of 7,421 instances, resulting in 50,537 instances across all sets.

The TRMF dataset significantly surpasses both MATRES and TCR in the number of annotated instances, offering a more extensive and diverse collection of temporal event relationships. TRMF's richness establishes it as an invaluable resource for research in this field. It enables more robust training and evaluation of models focused on understanding and predicting temporal relationships between events. Some interesting statistics will be presented in the appendix B.

4.2 Experimental Setup

In this study, we conducted experiments using the TRMF, MATRES, and TCR datasets. MATRES (Ning et al., 2018c) is a TempRel dataset that contains refined annotations on TimeBank (Cassidy et al., 2014) and TempEval (UzZaman et al., 2013) documents. TCR (Ning et al., 2018a) utilizes the same annotation scheme as MATRES. However, it contains a substantially smaller number of event-relation pairs.

We utilized RoBERTa-base as our pre-trained language model. Subsequently, we fine-tuned it for task-specific purposes to enhance its ability to extract event temporal relations. The learning rates for the pre-trained model and other parameters were set between {5e-6, 5e-5}. During virtual adversarial training, the noise variance was set to 1e-5, the noise scaling factor to 1e-6, and the adversarial step size to 1e-3. We conducted experiments using five different random seeds. We selected the learning rate and model with the best average performance on the development set for comparison on the test set. Our hyperparameter analysis is in Appendix C.

Baselines To validate the performance of the MFRV, we selected the following representative models for event temporal relation extraction for comparison:

Joint Constrained Learning (Wang et al., 2020) employs logical constraints to ensure holistic and logically consistent temporal relations.

HGRU (Tan et al., 2021) handles temporal relations using Hyperbolic Gated Recurrent Units and enhances the understanding of temporal relations with external knowledge.

Relative Event Time (Wen and Ji, 2021) links relative event time prediction with temporal relation extraction, incorporating relative event time as an additional feature in the classifier's training.

ChatGPT_ZS (Yuan et al., 2023) excels in handling zero-shot temporal relation extraction without needing prior task-specific training.

ChatGPT_ER (Yuan et al., 2023) benefits from a structured event ranking prompt that simplifies the task by asking for event orders.

ChatGPT_PE (Chan et al., 2024) incorporating manually designed task-specific prompts, guiding the model to focus on essential task features such as temporal order and event relations.

Bayesian-Trans (Tan et al., 2023) combines principles of statistical inference with deep learning methods. It utilizes Bayesian techniques to model the uncertainty in relation extraction.

PIPER (Zhang and Li, 2023) introduces a contrastive optimization framework that aligns model

Label Type	TRMF			MATRES			TCR		
	Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
BEFORE	74.9	51.5	61.1	**83.2**	**90.7**	**86.8**	**93.3**	**88.2**	**90.7**
AFTER	**76.4**	**92.7**	**83.7**	76.4	89.9	82.6	82.7	83.5	83.1
EQUAL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VAGUE	0.0	0.0	0.0	30.4	15.1	20.2	0.0	0.0	0.0
MICRO-AVERAGE	76.0	76.5	76.3	80.4	86.6	83.4	89.7	86.5	88.1

Table 2: Experimental Results for All Label Types on the Three Datasets.

outputs with logical constraints.

OntoEnhance (Zhuang et al., 2023a) improves event representation by integrating semantic information from event ontologies.

SDLG (Zhuang et al., 2023b) employs a syntax-based dynamic latent graph to enhance temporal relation extraction.

4.3 Experimental Results

Table 2 shows the experimental results for all label types by the MFRV across three datasets. MFRV achieved an F1 score of 76.3 on the TRMF dataset, 83.8 on the MATRES dataset, and 88.1 on the TCR dataset. These scores are the highest among the compared models, demonstrating the effectiveness of the MFRV in the task of event temporal relation extraction.

From Table 2, it can be observed that for the BEFORE and AFTER temporal relations, our model achieved relatively high Precision and Recall values. Notably, for the AFTER relation, the Recall values reached an impressive 92.7%, 89.9%, and 83.5%, respectively. These results indicate that MFRV is capable of identifying the majority of AFTER relations. For the BEFORE relation, the Precision values reached an impressive 74.9%, 83.2%, and 93.3%, respectively. These results indicate that MFRV can accurately identify BEFORE relations. In terms of Micro-average, MFRV attained over 76% in Precision, Recall, and F1 scores, demonstrating its overall strong performance.

Table 3 lists the performance comparison between the MFRV and the baseline models on the TRMF dataset. It can be seen that our MFRV achieved Precision, Recall, and F1 scores of 76.0%, 76.5%, and 76.3%, respectively, surpassing the aforementioned baseline models in overall metrics. These results demonstrate the effectiveness of our model in extracting temporal relations of events in the military field.

Table 4 lists the performance comparison between the MFRV and baseline models on the MATRES dataset. The MFRV achieved the highest F1 score, surpassing the previously best SDLG model by 0.3 percentage points. This demonstrates the effectiveness of the MFRV in event temporal relation extraction in a general field. The MFRV also achieved the highest recall rate, indicating its ability to comprehensively identify various positive instances and variations. Performance Comparison with Baseline Models on the TCR Dataset will be presented in the appendix D.

From the experimental results, it can be seen that the current LLM still has a significant gap compared to supervised methods in the field of event temporal relation extraction. However, we believe in the great potential of LLM, and we plan to conduct research combining LLM in the future.

Error Analysis and Discussion We conducted an error analysis on the experimental results across

Model	Precision	Recall	F_1
HGRU (Tan et al., 2021)	64.1	70.2	67.1
Relative Event Time (Wen and Ji, 2021)	73.4	**77.0**	75.1
Bayesian-Trans (Tan et al., 2023)	71.2	72.0	71.6
ChatGPT_ZS (Yuan et al., 2023)	40.0	37.3	38.6
ChatGPT_ER (Yuan et al., 2023)	38.1	33.2	35.5
ChatGPT_PE (Chan et al., 2024)	48.2	46.1	47.1
MFRV	**76.0**	76.5	**76.3**

Table 3: Performance Comparison with Baseline Models on the TRMF Dataset.

Model	Precision	Recall	F_1
Bayesian-Trans (Tan et al., 2023)	79.6	86.0	82.7
PIPER (Zhang and Li, 2023)	-	-	81.8
OntoEnhance (Zhuang et al., 2023a)	79.0	86.5	82.6
SDLG (Zhuang et al., 2023b)	**82.0**	84.2	83.1
ChatGPT_ZS (Yuan et al., 2023)	26.4	24.3	25.3
ChatGPT_ER (Yuan et al., 2023)	21.9	17.3	19.3
ChatGPT_PE (Chan et al., 2024)	-	-	35.0
MFRV	80.4	**86.6**	**83.4**

Table 4: Performance Comparison with Baseline Models on the MATRES Dataset.

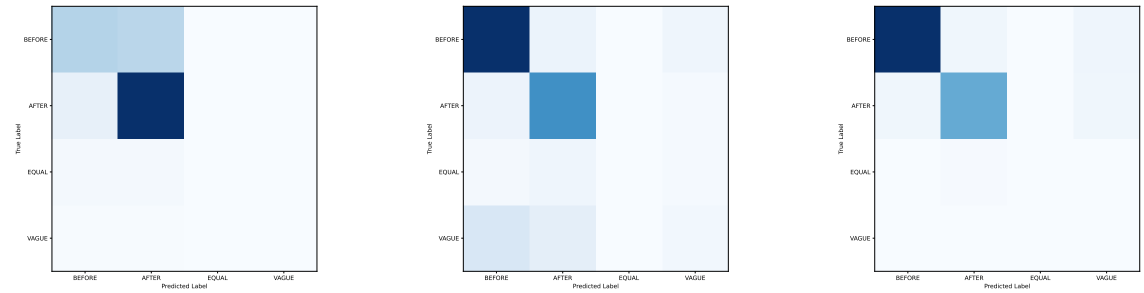


Figure 4: Confusion Matrix of Final Results on Three Datasets.

the three datasets. Figure 4 presents the confusion matrix of the final results on three datasets. The color intensity in these figures reflects the number of samples in different regions, with darker colors indicating a higher number of samples.

On the TRMF dataset, the confusion matrix reveals that many BEFORE samples were misclassified as AFTER. Similarly, among the actual AFTER samples, 342 were misclassified as BEFORE, while 4317 were correctly classified. This type of error indicates that the model has some difficulty distinguishing between BEFORE and AFTER relations. This may be due to the similar linguistic features of these two relations in context, leading to model confusion. From the third and fourth rows of the matrix, it is evident that the model fails to correctly classify EQUAL and VAGUE relations. This results in EQUAL and VAGUE samples being misclassified as either BEFORE or AFTER. We believe this error is due to the low number of EQUAL and VAGUE samples in the dataset, which prevents the model from adequately learning the characteristics of these relations during training.

An error example illustrates this challenge: *"The broadcasts were prerecorded (E1) two days before, suggesting the evacuation (E2), renewed shelling (E3), and other events, including an inexplicable car bombing (E4), in Donbas are being orchestrated by the Kremlin, say (E5) Ukrainian officials."*

The true temporal order is (E1 $\rightarrow$ E2 = E3 = E4 $\rightarrow$ E5), while the model predicted (E1 $\rightarrow$ E2 $\rightarrow$ E3 $\rightarrow$ E4 $\rightarrow$ E5). This example demonstrates that our model tends to confuse the order of events when dealing with multiple overlapping events, often identifying them as sequential BEFORE and AFTER relations. This may be due to the model's reliance on relative time prediction. In future research, we plan to design a module for global timeline construction to better capture and reflect complex event relationships. Additionally, recognizing EQUAL and VAGUE relationships can be more subtle and context-dependent than BEFORE and AFTER relationships, resulting in lower accuracy. We plan to introduce more finely annotated training data in future research. Error Analysis and Discussion on the MTRES and TCR dataset will be presented in the appendix E.

Model	TRMF	MATRES	TCR
Ours (BERT)	75.1	83.0	86.9
Ours (RoBERTa)	**76.3**	**83.4**	**88.1**
Ours (RoBERTa) w/o TEMPROB	75.6	82.6	86.9
Ours (RoBERTa) w/o RT	71.0	82.8	86.6
Ours (RoBERTa) w/o VAT	75.4	82.3	85.4

Table 5: Ablation Study of the MFRV on Three Datasets.

Ablation Study we designed an ablation study to understand the importance of RoBERTa, relative event time prediction, and virtual adversarial training combined with the MoCo method. The experimental results are shown in Table 5. To better explore the impact of word embedding representations on the performance of the MFRV model, we replaced RoBERTa with BERT (Devlin et al., 2019). As shown in the table, RoBERTa provides better performance compared to BERT. In Table 5, RT represents relative time prediction, and VAT represents virtual adversarial training.

It is evident from the table that both relative event time prediction, TEMPROB knowledge base, and virtual adversarial training combined with the MoCo method improve the model's performance. Removing either component results in a loss of performance, which strongly demonstrates the effectiveness of our proposed methods.

5 Conclusion

We propose an event temporal relation extraction model, MFRV, designed for the military field, based on relative event time prediction and virtual adversarial training. Relative event time prediction serves as an auxiliary task to enhance the model's understanding of event temporal relationships, thereby optimizing its performance. Virtual adversarial training aims to enhance the model's generalization capabilities by generating adversarial samples. We combine virtual adversarial training with the MoCo method, thereby improving the model's convergence and overall performance. To address the lack of datasets for event temporal relation extraction in the military field, we constructed a dataset named TRMF through a combination of large language models and human review. Experimental results on the TRMF, MATRES, and TCR datasets demonstrate the effectiveness of the MFRV in event temporal relation extraction tasks in both military and general fields.

Limitations

We acknowledge that the datasets we create and utilize still exhibit certain limitations in terms of quality and scale. In the research on event temporal relation extraction, the issues of dataset quality and scale have consistently posed significant challenges. Considering the high standards that deep learning systems demand for data, these problems are particularly pronounced. A promising future direction is to enhance both the quality and scale of datasets, which will not only enrich data resources but also improve the effectiveness and precision of deep learning techniques. Additionally, modeling global timelines is crucial to addressing the issue of error propagation in practical applications. An incorrect event timeline can disrupt the causal chain of events, which may lead to faulty decision-making and misinterpretation of event relationships. By accurately modeling the overall event timeline, we can minimize such errors and ensure more reliable outcomes. Our current work focuses solely on temporal relations between events. In the future, this work could be extended to include other types of event relationships, such as causal relations, entailment relations, and more.

Ethical Considerations

The objective of the proposed method is to identify and determine the temporal relationships between different events in a text, including establishing the sequence in which events occur. In the most optimistic scenario, the method can achieve results comparable to providing the same text to a human reader and having them explain the event relations. Therefore, ethical considerations are primarily focused on the construction of the dataset. In this paper, the raw data used to construct our dataset is sourced from publicly available news articles. As long as users employ the data legally, our dataset and methods will not have any direct harmful impact. However, we emphasize the importance of using the TRMF dataset in a responsible manner. Users should avoid applications that may exacerbate conflicts or cause harm to civilians. The dataset should be utilized for constructive purposes, such as crisis management or humanitarian efforts, to mitigate the impact of disasters and improve decision-making processes in sensitive situations. Ethical usage of the dataset is critical to ensure it contributes positively to societal needs without inadvertently fueling harmful actions.

References

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5412–5417, Online. Association for Computational Linguistics.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 189–198, Sydney, Australia. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Fei Cheng, Masayuki Asahara, Ichiro Kobayashi, and Sadao Kurohashi. 2020. Dynamically updating event representations for temporal relation classification with multi-category learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1352–1357, Online. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1796–1802, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Heshan Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. 2023. Mitigating gradient bias in multi-objective learning: A provably convergent approach. International Conference on Learning Representations.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019a. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019b. Joint event and temporal relation extraction with shared representations and structured prediction. 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 434–444, Hong Kong, China. Association for Computational Linguistics.

EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. 2022. 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.

Timotej Knez and Slavko Žitnik. 2024. Multimodal learning for temporal relation extraction in clinical texts. *Journal of the American Medical Informatics Association*, 31(6):1380–1387.

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246, Brussels, Belgium. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

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*.

Yuanliang Meng and Anna Rumshisky. 2018. Context-aware neural model for temporal information extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 527–536, Melbourne, Australia. Association for Computational Linguistics.

Yuanliang Meng, Anna Rumshisky, and Alexey Romanov. 2017. Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Copenhagen, Denmark. Association for Computational Linguistics.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. 2018. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. 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.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018b. Improving temporal relation extraction with a globally acquired statistical resource. 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 841–851, New Orleans, Louisiana. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018c. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

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

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2023. Event temporal relation extraction with Bayesian translational model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1125–1138, Dubrovnik, Croatia. Association for Computational Linguistics.

Xingwei Tan, Yuxiang Zhou, Gabriele Pergola, and Yulan He. 2024. Set-aligning framework for auto-regressive event temporal graph generation. In *Proceedings of 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second joint conference on lexical and computational semantics (* SEM), volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 1–9.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

Haoyang Wen and Heng Ji. 2021. Utilizing relative event time to enhance event-event temporal relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10431–10437, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. 2009. Jointly identifying temporal relations with Markov Logic. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 405–413, Suntec, Singapore. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. 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.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2024. Temporal relation extraction with contrastive prototypical sampling. *Knowledge-Based Systems*, 286:111410.

Beibei Zhang and Lishuang Li. 2023. Piper: A logic-driven deep contrastive optimization pipeline for event temporal reasoning. *Neural Networks*, 164:186–202.

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.

Youwen Zhao, Xiangbo Yuan, Ye Yuan, Shaoxiong Deng, and Jun Quan. 2023. Relation extraction: advancements through deep learning and entity-related features. *Social Network Analysis and Mining*, 13(1):92.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022. RSGT: Relational structure guided temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2001–2010, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ling Zhuang, Hao Fei, and Po Hu. 2023a. Knowledge-enhanced event relation extraction via event ontology prompt. *Information Fusion*, 100:101919.

Ling Zhuang, Hao Fei, and Po Hu. 2023b. Syntax-based dynamic latent graph for event relation extraction. *Information Processing & Management*, 60(5):103469.

A Prompts and Example

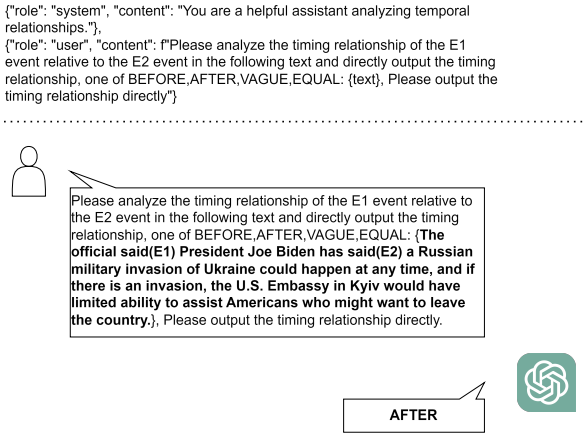


Figure 5: Our Utilized Prompts and Example.

B The statistical data of TRMF

Figure 6 presents the top 10 cities and countries where military news events occurred most frequently. The top four cities are Washington, Moscow, Kyiv, and London. This result aligns with common sense since national leaders typically announce political views or decisions from their offices in the capitals. Correspondingly, the four countries that appear most frequently are the United States, Russia, Ukraine, and the United Kingdom. As shown in Figure 7, these four major countries are the most closely associated with military issues related to Ukraine.

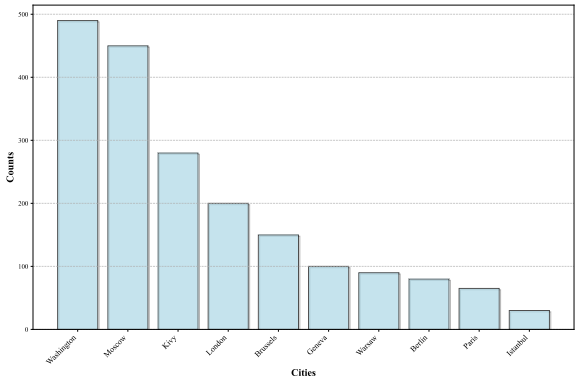


Figure 6: Top Ten Cities in Military News Locations.

C Hyperparameter Analysis

We conducted a grid search to determine the optimal hyperparameter combination for the model. The search range for the learning rate was 1e-6, 5e-6, 1e-5, 5e-5, 1e-4, for the noise variance was 1e-7, 1e-6, 5e-6, 1e-5, 1e-4, for the noise scaling factor was 1e-8, 1e-7, 5e-7, 1e-6, 5e-6, 1e-5, and for the adversarial step size was 1e-4, 5e-4, 1e-3, 5e-3, 1e-2. The experimental results indicate that hyperparameter tuning has a certain impact on model performance, but overall, the performance remains stable.

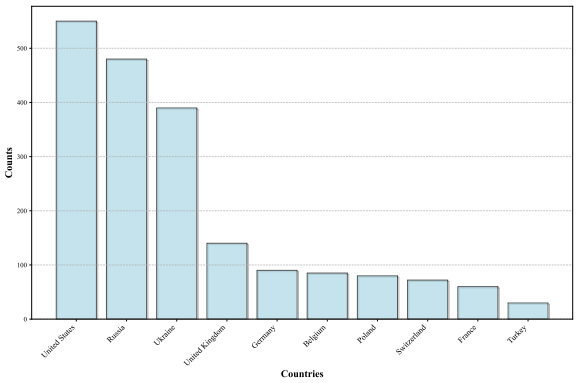


Figure 7: Top Ten Countries in Military News Locations.

D Performance Comparison with Baseline Models on the TCR Dataset

Table A1 lists the performance comparison between the MFRV and baseline models on the TCR dataset. The MFRV achieved the highest F1 score, surpassing the previously best OntoEnhance model by 1.3 percentage points. The TCR dataset primarily focuses on two types of temporal relations: BEFORE and AFTER. The MFRV achieved an F1 score of 90.7 in the BEFORE category and an F1 score of 83.1 in the AFTER category, which demonstrates the effectiveness of the MFRV on the TCR dataset.

Model	Precision	Recall	F_1
Joint Constrained Learning (Wang et al., 2020)	83.9	83.4	83.7
Bayesian-Trans (Tan et al., 2023)	**89.8**	82.6	86.1
OntoEnhance (Zhuang et al., 2023a)	89.6	84.3	86.8
ChatGPT_ZS (Yuan et al., 2023)	29.2	27.5	28.3
ChatGPT_ER (Yuan et al., 2023)	25.7	21.7	23.5
ChatGPT_PE (Chan et al., 2024)	37.9	36.8	37.3
MFRV	89.7	**86.5**	**88.1**

Table A1: Performance Comparison with Baseline Models on the TCR Dataset.

When comparing the performance of the MFRV across three datasets, we found that the model exhibited a distinct overall advantage on the TCR dataset. Upon further analysis, we discovered that the TCR dataset contains a higher proportion of BEFORE and AFTER temporal relations. Given that

the MFRV is designed with a relative event time prediction capability, it tends to favor predicting relations as BEFORE and AFTER. Consequently, this predisposition results in better performance on the TCR dataset.

E Error Analysis and Discussion

On the MATRES dataset, the model is relatively more successful in distinguishing between BEFORE and AFTER relations, though there is still some confusion between these two categories. We hypothesize that this may be due to potential variations in the quality and consistency of the annotation within the training data. It is also observed that the VAGUE category significantly impacts the model's judgments, particularly when the sample size is relatively small. Some samples that originally belonged to the BEFORE and AFTER categories are predicted as VAGUE, and conversely, some VAGUE samples are incorrectly classified as BEFORE and AFTER. We believe this may be due to the imbalance in the number of samples between types, which has affected the model's learning and generalization capabilities. Compared to this, the recognition errors between the BEFORE and AFTER categories are fewer. This suggests that improving the accuracy of determining the presence of temporal relations can effectively enhance model performance.

On the TCR dataset, we found that the model struggles to effectively identify the EQUAL category. Samples originally belonging to the EQUAL category are misclassified as other categories by the model. We believe this is due to the limited number of EQUAL samples in the training data, which restricts the model's ability to learn this relationship during training and results in poor generalization capability. Therefore, addressing the issue of small sample sizes in categories like EQUAL is a significant direction to consider in future work.