Temporal Cognitive Tree: A Hierarchical Modeling Approach for Event Temporal Relation Extraction

Anonymous ACL submission

Abstract

Understanding and analyzing event temporal relations is a crucial task in Natural Language Processing (NLP). This task, known as Event Temporal Relation Extraction (ETRE), aims to identify and extract temporal connections between events in text. Recent studies focus on locating the relative position of event pairs on the timeline by designing logical expressions or auxiliary tasks to predict their temporal occurrence. Despite these advances, this modeling approach neglects the multidimensional information in temporal relation and the hierarchical process of reasoning. In this study, we propose a novel hierarchical modeling approach for this task by introducing a Temporal Cognitive Tree (TCT) that mimics human logical reasoning. Additionally, we also design a integrated model incorporating prompt optimization and deductive reasoning to exploit multidimensional supervised information. Extensive experiments on TB-Dense and MATRES datasets demonstrate that our approach outperforms existing methods.

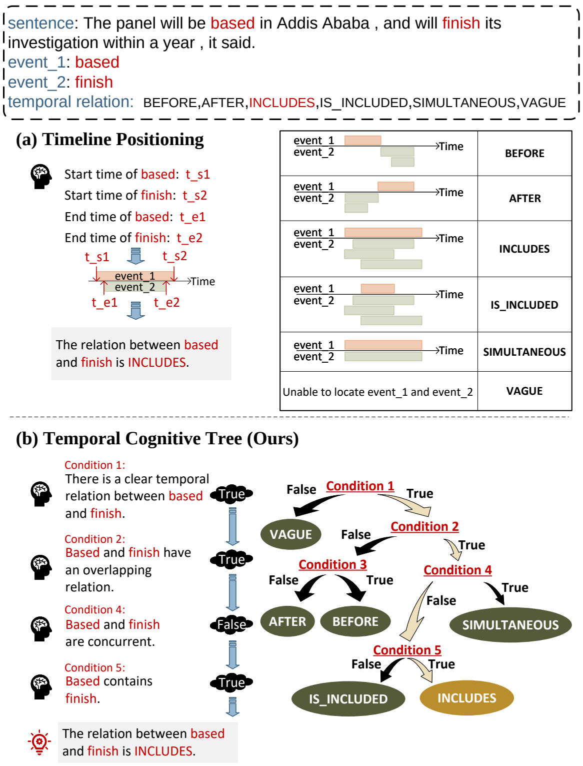

Figure 1: An example of ETRE task and two different modeling methods.

1 Introduction

Event relations usually refer to the mutual connections and influences between events. Understanding and analyzing event relations are crucial for individuals to comprehend the world. In the field of Natural Language Processing (NLP), extracting temporal relations between events is a critical task that aims to identify and interpret the temporal connections within textual data, as illustrated in Figure 1, given a sentence containing two events and a set of candidate temporal relations, our objective is to determine that the relation between the Event1 **based** and the Event2 **finish** is *INCLUDES*.

Researchers have invested substantial effort in the Event Temporal Relation Extraction (ETRE) task and have explored this topic in various ways. Early work primarily relied on traditional machine learning and statistical methods (Mani et al., 2006; Yoshikawa et al., 2009; Fei et al., 2020). In recent years, many studies have attempted to incorporate external knowledge to alleviate the issue of data scarcity in ETRE. Extensive experiments have demonstrated that augmenting knowledge can enhance model performance (Ning et al., 2019; Wang et al., 2020; Han et al., 2020; Tan et al., 2023; Zhuang et al., 2023). However, relying on external knowledge inevitably brings new challenges, such as noise injection and the model's over-reliance on external knowledge. Furthermore, recent studies have emphasized the importance of temporal relation semantics, treating it not merely as a conventional multi-class classification task but rather focusing on the relative positions of events on the timeline (Leeuwenberg and Moens, 2018; Wen and

Ji, 2021; Huang et al., 2023). However, existing methods based on timeline positioning only utilize the occurrence times of events to infer temporal relations, as illustrated in Figure 1(a). This modeling approach can merely consider the semantics of temporal relations linearly, i.e., the determination of temporal relations depends simply on a linear combination of start and end times of event pairs, which overlooks the hierarchical transitivity inherent in the process of reasoning. Consequently, the model can simply learn limited information about the position of events on the timeline from single-dimensional information, and fails to learn more multidimensional semantic knowledge, which may lead to the model's lack of understanding of temporal relations, such as the *VAGUE* relation, its complex semantic meaning can easily cause the model to misclassify other relations as *VAGUE*.

To enable the model to fully leverage the hierarchical prior knowledge in the process of inference, and thus learn the intrinsic meaning of temporal relations from multiple dimensions, we model the task of ETRE in a hierarchical manner and propose a ETRE model that integrates prompt optimization and deductive reasoning. To be specific, we design a Temporal Cognitive Tree (TCT), as illustrated in Figure 1(b), which is more consistent with human thinking patterns. Based on the TCT, we propose two modules, firstly, in order for the model to fully leverage the multidimensional supervised information in the TCT for training, we design a temporal relation judgment module based on multi-task prompt optimization. Secondly, to better leverage hierarchical information in the reasoning process, we propose a temporal inference module based on deductive reasoning. Extensive experiments demonstrate that our method can help the model better recognize the temporal relations between events.

Our contributions can be summarized as follows:

- We propose a novel approach to hierarchically model the existing task of ETRE by presenting a Temporal Cognitive Tree based on human logical reasoning. On the basis of this cognitive tree, we design a temporal relation extraction model that integrates prompt optimization and deductive reasoning.

- We present a multi-task temporal relation judgment module based on prompt optimization, and a multi-label temporal relation inference module based on deductive reasoning.These two modules leverage multidimensional knowledge in the hierarchical reasoning process to assist the model in better discerning the temporal relations between event pairs.

- We evaluate our model on two publicly available datasets, TB-Dense and MATRES. Experimental results demonstrate that our approach achieves state-of-the-art (SOTA) performance without relying on external knowledge.

2 Method

In this section, we will introduce our entire model. Our overall model is illustrated in Figure 2. First, we will define the task of event temporal relation extraction. Then, we will present the design of our Temporal Cognitive Tree (TCT). Following this, we will present two modules proposed in our model based on TCT: a temporal judgment module based on multi-task prompt optimization, and a temporal inference module based on deductive reasoning. Finally, we will explain how we integrate these two modules to obtain the final temporal relation extraction model.

2.1 Problem Formulation

Given a sentence and the two events it contains, our objective is to determine the temporal relation between these two events. This task is typically regarded as a text classification task. The model's input generally includes a text segment and two event trigger words within this text for which the temporal relation needs to be determined. The output is a label that signifies a particular temporal.

2.2 Temporal Cognitive Tree

In different temporal relation extraction datasets, the number and meaning of temporal relations are different. In the TB-Dense dataset, temporal relations are defined in a fine-grained manner, for example, a *BEFORE* relation between event pairs (e_1, e_2) requires meeting the following two conditions simultaneously: a) e_1 starts earlier than e_2; b) e_1 and e_2 do not overlap on the timeline. However, in the MATRES dataset, determining a *BEFORE* relation between event pairs does not require condition b). Due to the variations in the methods of defining temporal relations, we design different temporal cognitive trees, as shown in Figure 3. These trees consist of two components: conditional prompts and a multi-label mapping rule.

Specifically, for each data point in a dataset with k types of temporal relations, we do not directly

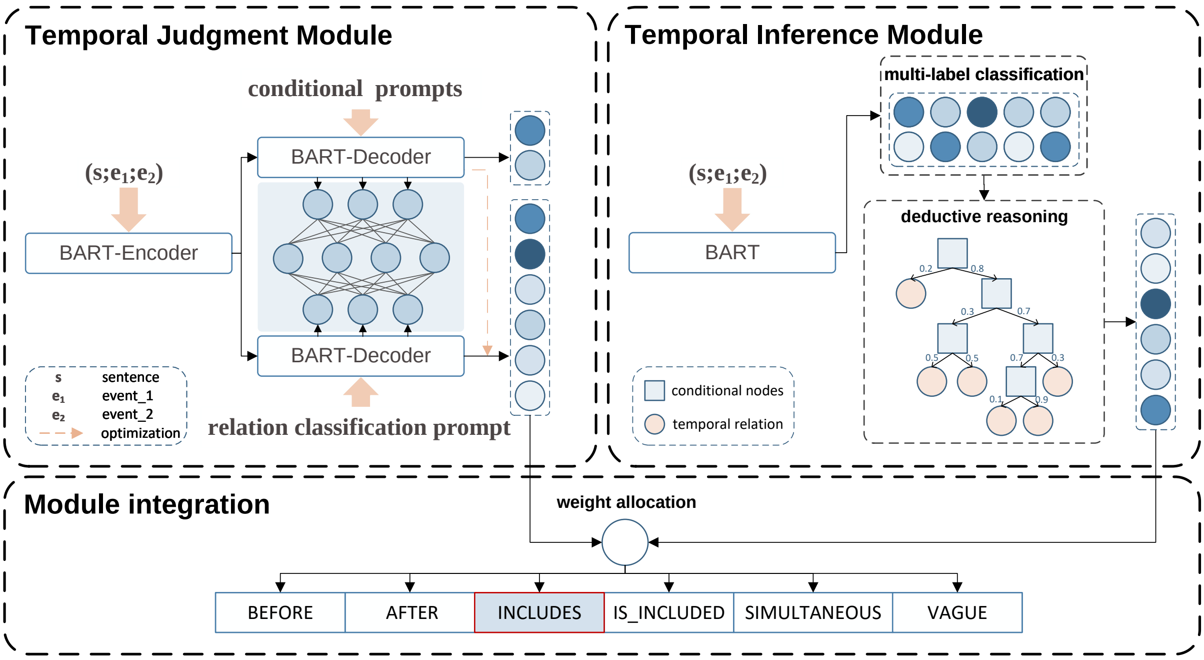

Figure 2: An overview of our model architecture.

inquire about the temporal relation of the given event pairs. Instead, we address the characteristics of temporal relations by asking yes or no questions from $k-1$ dimensions, thereby obtaining hierarchical temporal judgment information. For each question, we denote the answer "Yes" as label 1 and "No" as label 0. Each temporal relation can then be represented as a combination of $k-1$ binary values (0 and 1), resulting in a multi-label corresponding to each temporal category.

The temporal cognitive tree classifies each temporal relation in a fine-grained manner from different dimensions, thus transforming the original single-label problem into a multi-label problem. In addition to ensuring that all combinations of $0-1$ vectors for temporal categories are linearly independent, we design the cognitive tree based on the following two principles:

A) There should be consistency between different temporal categories in at least one dimension. We avoid designing multidimensional labels that are merely one-hot encodings of the original labels. Instead, we aim for the designed rules to help the model learn that different temporal categories share the same feature in at least one dimension, thereby facilitating a better comprehension of the temporal categories' meanings and the finer-grained differences.

B) All dimensions of any temporal category should be hierarchical. We intend for the designed prompt to present a process similar to human judgment of temporal relations, where higher-level judgment information is more abstract, and lower-level judgment information is more concrete. The labels of high-level prompts can determine the content of low-level prompts, and for some temporal categories, not all prompts needs to be used to determine them.

According to the principle **B)**, we find that we only need to ask certain higher-level judgment questions about event pairs to infer their temporal relations. Consequently, we can summarize the reasoning paths based on conditional prompts for temporal labels, as shown in the Table 1, where we use logical expressions to describe the reasoning paths. In Section 2.4, we will utilize these reasoning paths for temporal relation inference.

2.3 Temporal Judgment Module Based on Multi-Task Prompt Optimization

Our goal is to train a language model that can comprehend and determine the temporal relations between pairs of events accurately. It is obvious that according to our proposed cognitive tree, a robust language model should not only be capable of judging the temporal relation of (e_1, e_2) correctly, but also provide proper answers to the questions in the cognitive tree. We argue that additional training of the model to understand the semantic correlations and differences among the relations from different

Conditional Prompts	Multi-label Mapping Rule					
TB-Dense	BEFORE	AFTER	INCLUDES	IS_INCLUDED	SIMULTANEOUS	VAGUE
1. Is there a clear temporal relation between Event1 and Event2?	1	1	1	1	1	0
2. Do Event1 and Event2 have an overlapping relation?	0	0	1	1	1	0
3. Does Event1 precede Event2?	1	0	0	0	0	0
4. Are Event1 and Event2 concurrent?	0	0	0	0	1	0
5. Does Event1 contain Event2?	0	0	1	0	0	0
MATRES	BEFORE	AFTER			EQUAL	VAGUE
1. Do Event1 and Event2 occur in a clear and unique sequence?	1	1			0	0
2. Are Event1 and Event2 simultaneous?	0	0			1	0
3. Does Event1 precede Event2?	1	0			0	0

Temporal Cognitive Tree

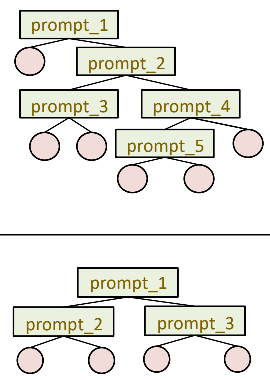

Figure 3: Details of the temporal cognitive trees corresponding to different manners of defining temporal relations.

perspectives is essential, which can help to make the language model better at discerning the temporal relations between event pairs.

We use a sequence-to-sequence model as the backbone architecture. We consider judging the conditional judgment prompts in the cognitive tree as the auxiliary task, while the determination of temporal relations between event pairs as the main task, and the model is trained in a multi-task manner. Specifically, we format the data into $(s; e_1; e_2)$, where s represents the sentence containing two events, and e_1 and e_2 represent the event pair for which the temporal relation needs to be determined. We take $x = (s; e_1; e_2)$ as the input for the model, and we extract the last layer's hidden state from the encoder part as the text encoding, which will be served as part of the input to the decoder.

After obtaining the text encoding, we interact it with the conditional prompts to obtain sentence representations that entail the hierarchical information. To be specific, for data with t temporal categories, we denote the conditional prompts as p_1, p_2, ..., p_{t-1}, and the final temporal relation classification prompt as f. In the decoder part, we input the conditional prompt list $[p_1, p_2, \ldots, p_{t-1}]$ along with the text encoding into the model sequentially. During the decoding process, the text encoding interacts with each token in the prompt text and obtains the special end-of-sequence token <eos> at the end of the prompt text as the final sentence representation h. Consequently, we can obtain a list of sentence representations $[h_p, h_f]$=$[h_1, h_2, ..., h_{t-1}, h_f]$ yielded from the interaction between each conditional prompt and the text.

For the auxiliary task, we set up a binary classifier with the set of candidate binary labels denoted as $A = \{0, 1\}$. For each prompt information p_i, $i \in \{1, 2, \ldots, t-1\}$, we calculate the loss $\mathcal{L}_i$ based on its corresponding binary label. Similarly, we define a multi-classifier as the final temporal relation classification layer for the main task, which we set the candidate labels as $M = \{r_1, r_2, \ldots, r_t\}$, representing the set of temporal relations, and compute the loss $\mathcal{L}_f$ according to the final temporal label. Therefore, we can construct the following two loss functions:

$$\mathcal{L}_i(\theta_{sh}, \theta_i) = \sum_{k=0}^{\|A\|} k \cdot \log(P_i(y = k \mid x)), \quad (1)$$

$$\mathcal{L}_f(\theta_{sh}, \theta_f) = \sum_{k=1}^{\|M\|} k \cdot \log(P_f(y = k \mid x)), \quad (2)$$

$$P_i(y = k \mid x) = softmax(\mathbf{MLP}_i(h_i)), \quad (3)$$

$$\begin{aligned} P_J(Y = r_k \mid x) &= P_f(y = k \mid x) \\ &= softmax(\mathbf{MLP}_f(h_f)), \end{aligned} \quad (4)$$

where y denotes the category number while Y denotes the final predicted temporal relation. θ_{sh} denotes the shared parameters for the main task and the auxiliary task, while θ_f and θ_i represent the remaining parameters for the main task and the auxiliary task during training respectively, excluding the shared parameters. $\mathbf{MLP}(\cdot)$ stands for task-specific multilayer perceptron.

We do not directly combine $\mathcal{L}_i$ and $\mathcal{L}_f$ through linear summation as the final training loss. Instead, inspired by the work of Sener and Koltun (2018), we treat the existing multi-task problem as a multi-objective optimization problem. We employ the

Dataset	Relation	Reasoning Path
TB-Dense	BEFORE	$P1 \wedge \neg P2 \wedge P3$
	AFTER	$P1 \wedge \neg P2 \wedge \neg P3$
	INCLUDES	$P1 \wedge P2 \wedge \neg P4 \wedge P5$
	IS_INCLUDED	$P1 \wedge P2 \wedge \neg P4 \wedge \neg P5$
	SIMULTANEOUS	$P1 \wedge P2 \wedge P4$
	VAGUE	$\neg P1$
MATRES	BEFORE	$P1 \wedge P3$
	AFTER	$P1 \wedge \neg P3$
	EQUAL	$\neg P1 \wedge P2$
	VAGUE	$\neg P1 \wedge \neg P2$

Table 1: The reasoning paths based on the temporal cognitive trees for different temporal relations. Here, *Pi* represents the i-th conditional information in the tree.

Multiple Gradient Descent Algorithm (MGDA) to search for the Pareto optimal solution in this task optimization process. For the optimization problem involving n auxiliary tasks and one primary task, we consider the parameters of the model's encoder as shared parameters, while the remaining parameters, i.e., those of the decoder and classification layers, are task-specific parameters. To achieve Pareto optimality, our multi-objective optimization problem is defined as follows:

$$\min_{\theta_{\mathrm{sh}},\theta_1,\ldots,\theta_{t-1},\theta_{\mathrm{f}}} \left(\mathcal{L}_1(\theta_{sh},\theta_1),\ldots,\mathcal{L}_f(\theta_{sh},\theta_f)\right)^{\mathrm{T}} \quad (5)$$

Following Sener and Koltun (2018), we transform the solution to Pareto optimality into a solution to task weights. We consider the optimization problem:

$$\min_{\alpha^1,\ldots,\alpha^{t-1},\alpha^f} \left\{ \left\| \sum_{i=1}^{T} \alpha^i \nabla_{\theta_{sh}} \mathcal{L}_i(\theta_{sh},\theta_i) \right\|_2^2 \right\}, \quad (6)$$

$$s.t. \sum_{i=1}^{T} \alpha^i = 1, \alpha^i \geq 0 \forall i, \quad (7)$$

where $T = \{1, 2, \ldots, t-1, f\}$, $\nabla_{\theta_{sh}} \mathcal{L}_i(\theta_{sh},\theta_i)$ is the gradient over the shared parameters.

Once the weights α^i is determined, the parameters θ_{sh} is updated using the weighted sum of the gradients:

$$\theta_{sh} = \theta_{sh} - \eta \sum_{i=1}^{T} \alpha^i \nabla_{\theta_{sh}} \mathcal{L}_i(\theta_{sh},\theta_i), \quad (8)$$

where η is the learning rate. θ_i updates in the normal way. The process is repeated for each iteration in the training, continually adjusting the parameters to move towards a Pareto optimal solution.

2.4 Temporal Inference Module Based on Deductive Reasoning

According to the TCT we designed, we argue that the determination of the temporal relation between any event pairs can be inferred based from a series of hierarchical prior knowledge ranging from abstract to concrete. Therefore, we conduct deductive reasoning on the judgment of each feature branch of the tree based on the model, thereby deriving the final temporal relation.

We first train the model to correctly classify the inference results at each node of the tree, then transform the task into a multi-label binary classification problem. Specifically, similar to the format described in Section 2.3, given a piece of text and its corresponding event pairs, we concatenate them as the input x for the BART model and obtain the text representation H. Additionally, for a dataset with t temporal relations, we define $F = \{d_1, d_2, \ldots, d_{t-1}\}$ as the set of hierarchical features, $C = \{0, 1\}$ as the set of possible values for each dimension of the features, the label for each dimension i is represented as y^i, $y^i \in C$. For the training of our model, in addition to utilizing Hamming loss, which is commonly used in multi-label classification tasks, we also apply focal loss (Lin et al., 2017) to our task, which is designed for training with imbalanced samples, to ensure more robust model training. Specifically, we calculate the loss $\mathcal{L}_{fc}$ as follows:

$$\begin{aligned}\mathcal{L}_{fc} = \sum_{i=1}^{\|F\|}\sum_{j=0}^{\|C\|} &\exp(\log\sigma(-logit_j^i(2y^i-1))\cdot\gamma) \\ &\cdot(logit_j^i\cdot(1-y^i)+mv+LSE(logit_j^i)),\end{aligned} \quad (9)$$

$$LSE(logit_j^i) = \log\left(e^{-mv} + e^{-logit_j^i - mv}\right), \quad (10)$$

where $mv = \max(-logit_j^i, 0)$ and $LSE(\cdot)$ means Log-Sum-Exp(LSE) operation, both of them are introduced to ensure numerical stability, γ acts as a modulation factor for the loss function, adjusting the contribution of different samples to the overall loss.

After training the model as described above, we obtain the classification probabilities for each event pair at the conditional nodes of the temporal cognition tree. We denote the probability that the value of the i-th feature is 1 as $Pr(Pi)$, which can be calculated as follows:

$$Pr(Pi) = sigmoid(\mathbf{MLP}_I(H)[i]), \quad (11)$$

we stipulate that when $Pr(Pi) > 0.5$, it can be concluded that the event labels the i-th feature as 1, which also indicates that it satisfies the condition *Pi*. Finally, we calculate the probability distribution for each temporal label and derive the final temporal relation prediction probability $P_I(Y = r_k \mid x)$ based on the reasoning rules in Table 1 and the following calculation rules:

$$\begin{aligned} P \wedge Q &= Pr(P) \cdot Pr(Q) \\ P \wedge \neg Q &= Pr(P) \cdot (1 - Pr(Q)), \end{aligned} \tag{12}$$

2.5 Method Integration

After obtaining the temporal label probability distributions from the aforementioned two modules, we perform a weighted summation of these two distributions to obtain the final temporal label probability distribution as follow:

$$P_{final}(Y = r_k \mid x) = \alpha \cdot P_J + \beta \cdot P_I \tag{13}$$

3 Experiments

3.1 Dataset

We conduct our experiments on two widely recognized datasets: TB-Dense (Cassidy et al., 2014) and MATRES (Ning et al., 2018), both of them are publicly available for temporal relation extraction task. TB-Dense is a dataset characterized by dense annotation for temporal relation extraction. It contains six types of relations: *BEFORE*, *AFTER*, *INCLUDES*, *IS_INCLUDED*, *SIMULTANEOUS*, and *VAGUE*. While MATRES is annotated using an innovative multi-axis annotation scheme that includes only four types of temporal relations: *BEFORE*, *AFTER*, *VAGUE* and *EQUAL*. In line with the latest work (Zhuang et al., 2023), we divide the dataset using the same manner as in previous studies (Wen and Ji, 2021; Han et al., 2019a).

3.2 Experimental Setup

Consistent with previous work (Han et al., 2019b), we use the micro-F1 score, excluding the *VAGUE* category, as the evaluation metric for both MATRES and TB-Dense. We compare our model with a series of representative works from the past three years, we categorized these comparison models into three groups: **1) Knowledge-augmented models:** These models incorporate external knowledge or additional training data during training through various methods(Cao et al., 2021; Tan et al., 2021, 2023; Zhuang et al., 2023). **2) Timeline positioning models:** These models utilize different techniques to directly or indirectly locate the relative position of events on the timeline(Wen and Ji, 2021; Huang et al., 2023). **3) Other benchmark models:** These methods do not fall into the above two categories but have demonstrated outstanding performance(Han et al., 2021; Hwang et al., 2022; Zhang et al., 2022). Additionally, we employ the generative model T5-large (Raffel et al., 2020) and BART-large (Lewis et al., 2019), which are also based on the encoder-decoder architecture, as two baseline model for comparison.

We use BART-large as our backbone model, and we employ Adafactor as the optimizer, with a learning rate warm-up ratio of 0.1. We set the batch size to 32. For TB-Dense, we set the learning rate to 3e-5, α to 0.19 and β to 0.81. For MATRES, we set the learning rate to 2e-5, α to 0.5 and β to 0.5. All experiments are trained for 50 epochs on the training set, and the model achieving the best performance on the validation set is selected as the final model for testing.

4 Results and Analysis

4.1 Overall Performance

As can be seen from the Table 2, without utilizing external knowledge, our proposed method consistently outperforms the existing methods and baseline models in the comparison of micro-F1. For the TB-Dense, our proposed method outperforms the existing SOTA method based on timeline positioning modeling by 2.9%, demonstrating the superiority of modeling the ETRE task based on TCT, which also indicates that compared to timeline position, the hierarchical knowledge in the TCT contains more information that is beneficial for model training. While for the MATRES, which only contains four types of temporal relations, despite the limited scale of the TCT we constructed (consisting of only three hierarchies) due to the nature of the temporal relations in MATRES, our novel approach outperforms the top result by a margin of 0.2%, showcasing the efficacy of TCT. Additionally, this also indicates that the greater the hierarchy of TCT, the higher the performance improvement in ETRE task, which highlights the importance of hierarchical information for model training. Furthermore, comparing with the two baseline models we constructed, we notice notable benefits of our suggested method on both TB-Dense and MA-

Model	Augmentation	TB-Dense			MATRES		
		P	R	F1	P	R	F1
Relative Time* (Wen and Ji, 2021)	-	-	-	-	78.4	85.2	81.7
Uncertainty-training (Cao et al., 2021)	✓	64.3	64.3	64.3	76.6	84.9	80.5
ECONET (Han et al., 2021)	-	-	-	66.8	-	-	79.3
HGRU (Tan et al., 2021)	✓	-	-	-	79.2	81.7	80.5
Probabilistic Box (Hwang et al., 2022)	-	-	-	-	-	-	71.1
Syntax Transformer (Zhang et al., 2022)	-	-	-	67.1	-	-	80.3
Bayesian-Trans (Tan et al., 2023)	✓	-	-	65.0	**79.6**	86.0	82.7
Unified-Framework* (Huang et al., 2023)	-	-	-	68.1	-	-	82.6
OntoEnhance (Zhuang et al., 2023)	✓	67.5	68.6	68.0	79.0	86.5	82.6
T5-large(Vanilla Classifier)	-	68.5	57.0	62.2	79.1	80.4	79.7
BART-large(Vanilla Classifier)	-	67.5	65.5	66.5	75.7	83.7	79.5
TCT(Ours)	-	**70.3**	**71.6**	**70.9**	79.0	**87.2**	**82.9**

Table 2: The overall experimental results on the TB-Dense and MATRES datasets. Models marked with a * use a timeline positioning modeling approach. Models with a check mark for "Augmentation" are knowledge-augmented models. All previous experimental results are cited from the data in their respective papers.

Dataset	Backbone	Method	P	R	F1
TB-Dense	BART-base	TCT	**66.8**	**62.7**	**64.7**
		w/o TJM	65.5	58.7	61.9
		w/o TIM	63.2	62.5	62.8
	BART-large	TCT	**70.3**	**71.6**	**70.9**
		w/o TJM	67.0	68.3	67.7
		w/o TIM	65.8	70.8	68.2
MATRES	BART-base	TCT	76.6	**82.7**	**79.5**
		w/o TJM	**76.8**	80.4	78.5
		w/o TIM	75.3	82.1	78.6
	BART-large	TCT	79.0	**87.2**	**82.9**
		w/o TJM	**79.3**	82.7	81.0
		w/o TIM	78.2	86.7	82.2

Table 3: The ablation experimental results on the TB-Dense and MATRES.

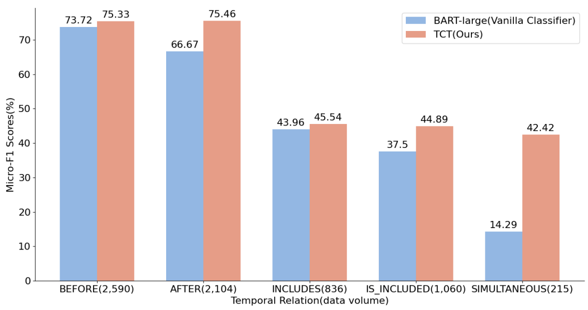

Figure 4: Comparison of micro-F1 values for each subcategory.

TRES, which further confirms the effectiveness of the TCT modeling approach.

4.2 Analysis of Results on Subcategories

We also analyze the classification results of our method on positive samples for each category in the TB-Dense. As shown in Figure 4, our method outperforms the baseline model in classifying each category, especially those with fewer instances, which indicates that our method can alleviate the impact of data imbalance on classification results to a certain extent. Furthermore, we compare the instances misclassified as *VAGUE* in the positive samples with the previous SOTA method, as shown in the Figure 5, which demonstrates a distinctive advantage in discerning ambiguous relation of our model.

4.3 Ablation Study

We conduct ablation experiments using two different sizes of backbone models (BART-base, BART-large). Based on the ablation study results shown in Table 3, we can draw the following conclusions:

1) Both the temporal judgment module (TJM) and the temporal inference module (TIM) have a non-ignorable impact on the overall model performance. For the TJM, in the TB-Dense, regardless of the model size, removing the TJM significantly reduces the overall model performance (by 2.8% and 3.2% respectively). Similarly, in the MATRES, removing the module also have a considerable impact on the overall model performance. For the TIM module, the experimental results in different sizes and datasets also demonstrate its significant effect on the overall performance. This illustrates the importance of utilizing multidimensional hierarchical semantic knowledge, which indeed facilitates the model to better identify the temporal relationships between events, and further demonstrates the effectiveness of the TCT modeling approach.

2) The fusion of the TJM and the TIM effectively combines their strengths. From the experimental re-

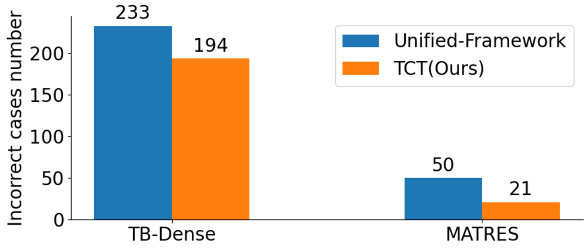

Figure 5: Comparison of the number of instances misclassified as relation *VAGUE*.

sults, it is evident that compared to TIM, TJM tends to improve the model's recall rate. Conversely, compared to TJM, TIM tends to achieve higher precision. This indicates that TJM is more advantageous in reducing erroneous predictions, while TIM is more beneficial in avoiding the omission of certain positive instances. The combination of these two modules naturally leverages their respective advantages, enabling the model to fully exploit its potential and achieve optimal performance.

4.4 Case Study

Figure 6 illustrates an example of our model in ETRE task. In this example, the model correctly identifies the relation between **finish** and **said** as *AFTER*, and notably, for each query within TCT, it provides accurate judgments. Clearly, this not only aligns with our expectations but also conforms to human common sense when assessing temporal relations. In addition, we show the value of the probability of the model's inference for each conditional branch in this example, which are available in the TIM. It is evident that the model's determination of the relation between **finish** and **said** as *AFTER* is based on its confident judgments for each conditional branch.

5 Related Work

Early works mainly utilized traditional machine learning and statistics-based methods for ETRE(Mani et al., 2006; Yoshikawa et al., 2009). With the development of deep learning, some works have combined pre-trained language models with graph-based models to improve encoding performance for alleviating the problem of long-distance dependency (Zhang et al., 2022; Mathur et al., 2021; Man et al., 2022). Some works focus on the problem of data scarcity in existing datasets, and propose to introduce external knowledge for knowledge enhancement (Ning et al., 2019; Wang et al., 2020; Han et al., 2020; Tan et al., 2023; Zhuang et al., 2023). There are also works that

Input: "Sentence: The panel will be based in Addis Ababa , and will finish its investigation within a year , it said. Event1: finish. Event2: said."
Q1: Is there a clear temporal relation between Event1 and Event2?
A1: 1 (Yes) **P(A1=1)=0.9987**
Q2: Do Event1 and Event2 have an overlapping relation?
A2: 0 (No) **P(A2=1)=0.0029**
Q3: Does Event1 precede Event2?
A3: 0 (No) **P(A3=1)=0.0045**
Q4: Are Event1 and Event2 concurrent?
A4: 0 (No) **P(A4=1)=0.0023**
Q5: Does Event1 contain Event2?
A5: 0 (No) **P(A5=1)=0.0019**
Q6: What's the temporal relation between Event1 and Event2?
Output: AFTER
P(relation=After)=0.9987*(1-0.0029)*(1-0.0045)=0.9913

Figure 6: An example of our model performing ETRE.

employ multi-task learning to compensate for the limitations of single-text classification tasks (Wen and Ji, 2021; Ballesteros et al., 2020; Cheng et al., 2020). Additionally, some of the latest work concerned with the significance of temporal semantics, and further enhanced the performance of temporal relation extraction by combining some rule constraints (Huang et al., 2023; Hwang et al., 2022).

Recently, the rapid development of Large Language Models(LLMs) has drawn attention to the potential of applying LLMs to ETRE task. Yuan et al. (2023) utilized prompt engineering techniques and conducted extensive experiments on ChatGPT to demonstrate that there is still considerable room for directly predicting on ChatGPT compared to supervised learning with smaller-scale models. Additionally, Huang et al. (2023) validated the limitations of ChatGPT in ETRE tasks in their work, with the best test result on the TB-Dense dataset achieving a micro-F1 score of 41.0%.

6 Conclusion and Future Work

In this paper, we propose a novel hierarchical modeling approach for ETRE. Specifically, we introduce a Temporal Cognitive Tree (TCT) that aligns with human logical reasoning processes. Our approach integrates prompt optimization and deductive reasoning, enhancing the model's ability to understand and extract temporal relations from a multidimensional perspective. Extensive experiments demonstrate that our approach achieves significant performance without the need for external knowledge. In future work, we aim to explore the possibilities of optimizing and extending this approach to accommodate relation extraction tasks with varying fields and data volumes.

Limitations

From an overall experimental result perspective, although our model outperforms the current SOTA results, it does not demonstrate an absolute advantage on the MATRES dataset (only 0.2% higher than the best result). We think this is due to our proposed method relying on the categories and quantity of temporal relations. Clearly, MATRES defines different temporal relations in a coarser granularity, resulting in fewer types of temporal relations, which limits the improvement potential of our method. Further research is needed to address the limitations of our proposed method in handling different quantities of temporal relations, in order to achieve a more robust model.

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.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. 2021. Uncertainty-aware self-training for semi-supervised event temporal relation extraction. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2900–2904.

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.

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.

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7692–7699.

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.

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.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. Econet: Effective continual pretraining of language models for event temporal reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5367–5380.

Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5717–5729.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9631–9646.

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.

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. *arXiv preprint arXiv:1808.09401*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11058–11066.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chungmin Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760.

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

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.

Qiang Ning, Hao Wu, and Dan Roth. 2018. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31.

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 (EMNLP)*, pages 8065–8077.

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.

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.

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 (EMNLP)*, pages 10431–10437.

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.

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.

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.

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