SCTc-TE: A Comprehensive Formulation and Benchmark for Temporal Event Forecasting

Yunshan Ma*
National University of Singapore
yunshan.ma@u.nus.edu

Chenchen Ye*
University of California, Los Angeles
ccye@cs.ucla.edu

Zijian Wu*
National University of Singapore
zijian.wu@u.nus.edu

Xiang Wang†
University of Science and Technology of China
xiangwang1223@gmail.com

Yixin Cao
Singapore Management University
caoyixin2011@gmail.com

Liang Pang
Institute of Computing Technology, Chinese Academy of Sciences
pangliang@ict.ac.cn

Tat-Seng Chua
National University of Singapore
dcscts@nus.edu.sg

ABSTRACT

Temporal complex event forecasting aims to predict the future events given the observed events from history. Most formulations of temporal complex event are unstructured or without extensive temporal information, resulting in inferior representations and limited forecasting capabilities. To bridge these gaps, we innovatively introduce the formulation of Structured, Complex, and Time-complete temporal event (SCTc-TE). Following this comprehensive formulation, we develop a fully automated pipeline and construct a large-scale dataset named MidEast-TE from about 0.6 million news articles. This dataset focuses on the cooperation and conflict events among countries mainly in the MidEast region from 2015 to 2022. Not limited to the dataset construction, more importantly, we advance the forecasting methods by discriminating the crucial roles of various contextual information, *i.e.,* local and global contexts. Thereby, we propose a novel method LoGo that is able to take advantage of both Local and Global contexts for SCTc-TE forecasting. We evaluate our proposed approach on both our proposed MidEast-TE dataset and the original GDELT-TE dataset. Experimental results demonstrate the effectiveness of our forecasting model LoGo. The code and datasets are released via https://github.com/yecchen/GDELT-ComplexEvent.

CCS CONCEPTS

• **Computing methodologies** → **Temporal reasoning**; • **Information systems** → **Specialized information retrieval**.

KEYWORDS

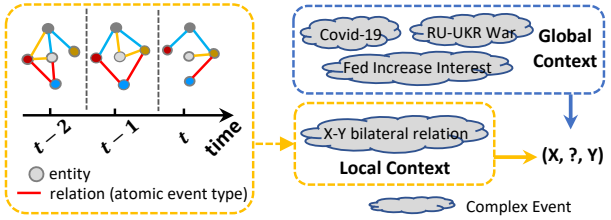
Temporal Event Forecasting, Temporal Complex Event, Temporal Knowledge Graph

1 INTRODUCTION

Temporal **E**vent (TE) forecasting conceptually targets at predicting a future event based on observed facts from history. People seek to mine the rules that govern the evolution of various events, in order to facilitate disaster prevention or early warning in various areas, such as civil unrests or regional conflicts. Due to its significant value, event forecasting has garnered more and more interests from the research communities.

Formulation	Datasets	Structured	Complex	Time-complete
Time Series	RCT-B	✗	✗	✓
Storyline	News14, WCEP	✗	✓	✓
TKG	GDELT, ICEWS	✓	✗	✓
TCE w Schema	General, IED	✓	✓	✗
SCTc-TE (ours)	MidEast-TE	✓	✓	✓

(a) Comparison between our SCTc-TE and previous formulations.



(b) Illustration of SCTc-TE. **(c) Effects of local/global contexts.**

Figure 1: (a) Comparison between our SCTc-TE and previous TE formulations; (b) An illustration of our SCTc-TE formulation; (c) The motivation of leveraging both local and global contexts.

Albeit the previous diverse formulations of TE forecasting, we highlight three key properties that are crucial for both representation and forecasting of TE, *i.e.,* **Structured**, **Complex**, and **Time-complete**. First, structured representations, such as Temporal Knowledge Graph (TKG) [3] or Temporal Complex Event (TCE) with schema [23], are simple and flexible for indexing and query, and the concise format of structured data enables the storing of large-scale events with limited memory. In contrast, other unstructured formulations, such as time series [31] or natural language [4], are either non-flexible to represent multi-line of events or redundant due to natural language representation. Second, Complex Event

*Equal contrication.
†Corresponding author. Xiang Wang is also affiliated with Institute of Artificial Intelligence, Institute of Dataspace, Hefei Comprehensive National Science Center.

(CE), which is composed by a set of atomic events [23], is capable of capturing multiple actors, relations, and timelines, thus satisfying various requirements in practice. Conversely, the formulations of atomic event, such as GDELT [22] or ACE2005 [41], are restricted to individual events that cannot model the complex scenarios or even perform forecasting. Third and more importantly, the time-complete characteristic, which means that every atomic event in a certain complex event has its corresponding timestamp, is indispensible. Some previous works [23] define a specific *temporal relation* to retain the temporal information within a complex event; however, it requires $O(N^2)$ temporal relations in order to fully preserve the temporal information of a complex event that has N atomic events. As a result, the temporal relations are rarely complete due to the quadratic relation space. In the summary as shown in Figure 1 (a), to the best of our knowledge, none of previous formulations satisfy all the three properties.

To bridge the gap, we borrow the merits of both TKG and TCE with schema and propose a novel formulation to represent the **S**tructured, **C**omplex, and **T**ime-**c**omplete **T**emproal **E**vent (SCTc-TE). Specifically, as shown in Figure 1 (b), we define one SCTc-TE as a list of semantically related and chronologically ordered graphs, each of which consists of a set of atomic events that are occurring at the same timestamp. We extend the atomic event formulation from the quadruple in GDELT [22] to a quintuple (s, r, o, t, c), where s, r, o, t and c represents the subject, relation [1], object, timestamp, and CE, respectively. In contrast to TKG, where all the atomic events are scattered, our formulation groups them into clusters, and each cluster corresponds to a CE. More importantly, each atomic event has an absolute timestamp, endowing our formulation with the time-complete property. In order to implement this new formulation, we develop a simple and fully automated pipeline, which utilizes the pre-trained large language models (LLMs) and time-aware clustering, to construct SCTc-TEs from news articles. Based on this pipeline, we construct two large-scale event forecasting datasets, MidEast-TE and GDELT-TE, from 0.6 million news articles.

Based on the newly proposed formulation and constructed datasets, we design a novel TE forecasting method that is able to leverage both local and global contexts. Given a particular query $(s, r, ?, t, c)$ to be forecast, we term the historical events belonging to the same complex event as the local context and all the historical events as the global context. The local context preserves the most relevant information that can be used to perform forecasting; while, the global context, which is usually noisy and scattered, provides a universal background that can also affect the evolution of certain complex events. For the example shown in Figure 1 (c), given the complex event that mainly depicts the bilateral relationship between countries X and Y, both the historical relations between these two countries and the global contexts, such as Covid-19 pandemic, RU-UKR war, and Federate Reserve increase interest, will affect the bilateral relations. However, previous approaches solely rely on either the global context [19, 26] or the local context [23], which is sub-optimal in forecasting complex events. In this work, we propose to unify the modeling of both local and global contexts for TE forecasting. Concretely, we adopt two context learning modules to separately learn the entity and relation representations under the local and global context, respectively. An early-fusion strategy followed by a decoder is then applied to achieve the final forecasting. Extensive experiments demonstrate that our model outperforms SOTA methods. The main contributions of this work are as follows:

- To the best of our knowledge, we are the first to propose the SCTc-TE formulation that encompasses all the structured, complex, and time-complete properties of TEs.
- Based on the SCT-TCE formulation, we develop a fully automated pipeline to construct TEs from news articles and construct two large-scale datasets, MidEast-TE and GDELT-TE.
- We propose a novel method (LoGo) that captures both the local and global contexts for SCTc-TE forecasting, and extensive evaluations and experiments demonstrate the richness of our datasets and the effectiveness of the proposed method.

2 PROBLEM FORMULATION AND DATASET

We first formally define the SCTc-TE and its forecasting task. Second, we develop a pipeline to construct SCTs-TE from news articles, as shown in Figure 2.

2.1 Problem Formulation

We define one SCTc-TE as a list of timestamped graphs $\mathbf{G}^c = [G_1^c, G_2^c, \cdots, G_t^c]$, where $c \in C$ is the identifier of one specific **Complex Event** (CE) and C is the entire identifier set for all CEs; and each graph is defined as $G_t^c = \{(s_n, r_n, o_n, t, c)\}_{n=1}^{N_t^c}$, where (s_n, r_n, o_n, t, c) is the n-th **atomic event** in G_t^c and N_t^c is the number of atomic events at timestamp t for the CE c. In terms of each atomic event, $s \in \mathcal{E}$, $r \in \mathcal{R}$, and $o \in \mathcal{E}$ correspond to the subject entity, relation, and object entity, respectively; t is the timestamp when the certain atomic event occurs; $\mathcal{E}$ and $\mathcal{R}$ are the entity and relation set, respectively. Typically, $\mathbf{G}^c \in \mathcal{G}$ is constructed based on a list of timestamped document set $\mathbf{D}^c = [D_1^c, D_2^c, \cdots, D_t^c] \in \mathcal{D}$, where $\mathbf{G}^c$ is the **local context** for all the atomic events that the CE c contains, while $\mathcal{G}$ is the **global context** that is a combined graph of all the CEs C. $\mathcal{D}$ is the entire document set that consists of a list of timestamped document set $[\mathcal{D}_1, \mathcal{D}_1, \cdots, \mathcal{D}_t]$. Note that, both D_t^c and $\mathcal{D}_t$ are a set of documents.

Given a set of document $\mathcal{D}$, the **SCTc-TE Construction** task aims to identify the CEs, *i.e.,* $\mathbf{D}^c$, from which we then extract the atomic events and form the SCTc-TE graph $\mathbf{G}^c$. Given a partial CE $\mathbf{G}_{\leq t}^c = [G_1^c, G_2^c, \cdots, G_t^c]$ and one query $(s, r, ?, t+1, c)$ at the next timestamp, the **SCTc-TE Forecasting** aims to predict the object o.

2.2 SCTc-TE Construction Pipeline

We first introduce the domain and data sources that we used to illustrate the pipeline. Following on, we explicate how to perform event extraction using LLMs, then we present how to identify CEs from news articles, and finally show the statistics and evaluations of our datasets.

2.2.1 Domain and Data Sources. We construct our datasets based on the GDELT [22] corpus, which is a large-scale TKG dataset that has publicly accessible URLs of news articles. More importantly, it follows a well-defined ontology, *i.e.,* CAMEO [2], which is authentic

[1] Here relation refers to the atomic event type.

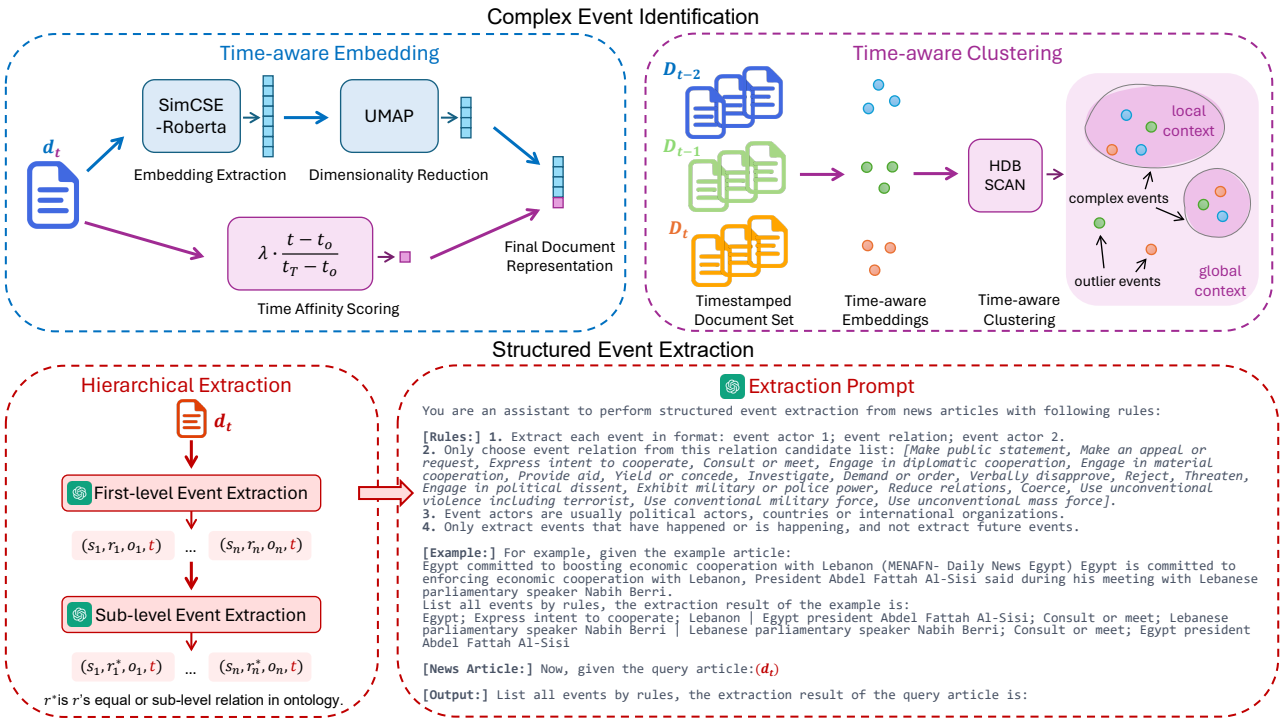


Figure 2: The dataset construction pipeline of SCTs-TE consists of: 1) *Complex Event Identification* clusters complex events by time-aware document embedding and 2) *Structured Event Extraction* hierarchically extracts structured events with LLMs.

in describing international political events and curated by well-known domain experts. As the original GDELT dataset is very large, we follow the previous approaches [8, 9] and crop a subset that is from three middle-east countries, *i.e.*, Egypt (EG), Iran (IR), and Israel (IS), for the period from 2015-02-19 to 2022-03-17. Then we download the event news articles with valid URLs, excluding those news articles of which the URLs are broken or inaccessible. To improve the article quality, we further applied document filtering steps by keeping popular and reliable news agencies and entities based on their frequency descending order. This step can largely reduce the amount of low-quality news articles, which are with poor writing or even fake news. Finally, we keep 275,406 documents which is about 47% of all successfully downloaded news articles.

2.2.2 Structured Event Extraction. Even though GDELT offers the extracted structured atomic events, unfortunately, such extraction results are prone to coarse-grained events and actors due to its outdated rule-based extraction system built more than ten years ago. Therefore, it is imperative to apply cutting-edge extraction techniques to re-extract the events. Most of the existing EE studies [27, 39, 41] are supervised methods, which require high-quality human-annotated datasets that are labor-intensive and expensive. With the striking success of ChatGPT [2] and GPT-4 [32], zero-shot information extraction without annotations becomes feasible [43, 49] which demonstrates high potential. We thus leverage LLMs for EE in a zero-shot paradigm. There are tens of optional commercial or open-source LLMs emerging and fast evolving, we take the well-performing open-sourced Vicuna-13b [3], because it is affordable for general academic labs *w.r.t.* computational resources and time considering million-level corpus.

Hierarchical Extraction. One major problem faced by LLM-based EE falls in the input length limitation, especially since the number of distinct atomic event types is over 200 and needs to be input together with necessary extraction instructions and source news articles. To solve this problem, we propose a hierarchical extraction pipeline based on the three-level relation hierarchy in CAEMO from coarse-grain to fine-grain. As shown in the bottom of Figure 2, we first input the news article and the first-level atomic event types, prompting the Vicuna to extract all the first-level events, each including event subject, object, and one of the first-level relations. We use the publish date of the news as the timestamp of the extracted events. Then, we parse the first-level extraction results. For each valid first-level event, we input its affiliated second-level relations as options to the Vicuna model, together with the original news article, thus obtaining the second-level extraction results. The same procedure applies to the third-level extraction. Note that we provide a '*No specific*' choice for the sub-level relation extraction, and in this case, the prior level relation will be kept.

Entity Linking. Since we do not have a pre-defined entity set during EE, the extracted entities have free forms and multiple entities correspond to the same one. For example, the extracted entities *U.S.A* and *United States* actually refer to the same actual entity. We

[2]https://chat.openai.com/

[3]https://chat.lmsys.org/ by May, 2023