Prompt-based Data Augmentation for Semantically-Precise Event Relation Classification

Youssra Rebboud[1,*], Pasquale Lisena[1] and Raphaël Troncy[1]

[1]*EURECOM, Sophia Antipolis, France*

Abstract

The process of recognizing and classifying the relationships between events mentioned in the text is a crucial task in natural language processing (NLP) known as event relation extraction. If temporal relations and causality are largely studied in the literature, other types of relations have found less interest. Our study specifically concentrates on four types of event relations: causality, enabling, prevention, and intention. Our main contribution consists of the use of a state-of-the-art language model (GPT-3) to extend an existing small dataset with synthetic examples to address the challenge of insufficient training data. We evaluate the quality of these generated samples by training an event relations extraction system, showing improved performances in classifying event relations.

Keywords

Event Relation, Information Extraction, Knowledge Graphs, Machine Learning

1. Introduction

Relation extraction (RE) – the identification and classification of relationships between two named entities in raw text – is a classic natural language processing (NLP) task which is receiving attention from the scientific community [1, 2]. A more specific branch of RE aims to automatically detect the relations between events (Event Relation Extraction, ERE). While, in RE, the entity type is crucial for inferring the correct relation[1], in ERE the relations involve homogeneous entities (namely, pairs of events), requiring specialised methods and approaches. Among event relation, the literature mostly focused on temporal relations [3], causality [4], and coreference of the same event in different textual resources [5].

Apart from causal and temporal structures, event relations may include different concepts such as prevention, intention, enabling, etc. Extracting this variety of relations may serve various downstream applications, including semantic timelines, question answering, and fact checking, supporting the decision making and improving information and entertainment. Apart

SEMMES: Semantic Methods for Events and Stories, May 28–23, 2023, Heraklion, Greece

[*]Corresponding author.

youssra.rebboud@eurecom.fr (Y. Rebboud); pasquale.lisena@eurecom.fr (P. Lisena); raphael.troncy@eurecom.fr (R. Troncy)

https://github.com/RYoussra (Y. Rebboud); http://pasqlisena.github.io/ (P. Lisena); https://www.eurecom.fr/~troncy/ (R. Troncy)

0000-0003-3507-5646 (Y. Rebboud); 0000-0003-3094-5585 (P. Lisena); 0000-0003-0457-1436 (R. Troncy)

[1]For example, the edge between “John” (Person) and “London” (City) can be “is resident in” and not “manufactured by” or “enrolled in”.

from a demonstrative proof-of-concept [6], the automatic extraction of different kind of relations have not been deeply investigated. The first step toward achieving this objective is to develop a dedicated dataset. Although [6] attempts to construct such an initial dataset for multiple event relations, the result is particularly small in size and exhibited significant imbalances.

The recent advent of generative models has marked a paradigm shift in Natural Language Processing with their capabilities to generate human-like complex text relying only on the provided prompt. In this work, we aim to use prompt-based solutions to extend *The Event Relations Dataset* [6] with synthetic sentences including references to event relations, with a particular focus on causality, prevention, intention, and enabling. This dataset will then be used to further evaluate state-of-art techniques based on machine learning, to check if they can be successfully applied in the prediction of event relation types.

This work has two research objectives:

- To investigate if prompt-based generative models are suitable for generating synthetic data for the purpose of populating a dataset of event relation, particularly for such kind of rare and very specific event relation types.
- Evaluate the performance of methods based on language models in predicting event relations when trained on synthetic data.

The remainder of this paper is structured as follows. First, we review the existing event relations datasets and event relations extraction in Section 2. We describe our approach for constructing a synthetic event relations dataset by proving GPT-3 with prompts to generate sentences that hold prevention, intention, and enabling relations as well as their constructs in Section 3. Section 4 details a method for extracting event relations from the generated dataset, whose results are discussed in Section 5. Finally, we conclude and outline some future work in Section 6.

2. Related Work

2.1. Events and Event Relationships Datasets

Within the field of events and events relationships, several datasets have been developed with the main objective of capturing events, event coreferences, causal and temporal relations. For instance, ACE 2005[2] for event extraction, and TimeBank [7] and CausalTimeBank [8] respectively for temporal and causal events relationship extraction.

In [6], the FARO ontology for representing event relations has been introduced, as a harmonisation of different models from the literature, including definitions for all relations. In addition, a first *Event Relation dataset* covering four event relation types: caus lity, prevention, enabli ng, and intention is presented in [6]. With the exception of causality, these relation types are absent in other annotated datasets. Due to the small size of the dataset and its imbalance, training a model on top of it presented a significant challenge.[3]

To address this gap, our work aims to increase the size of this dataset using data augmentation techniques. In prior research, numerous techniques were elaborated for data augmentation

[2] https://catalog.ldc.upenn.edu/LDC2006T06

[3] More details are provided in Section 3.1

within the field of events and events relationships extraction [5] such as distant supervision [9] and translation [10]. In this work, we intend to leverage capabilities of generative models such as GPT-3 [11]. It is worth to mention that, at the time of writing, no official API for ChatGPT is available.

2.2. Events and Events Relationship Extraction

Numerous techniques were adopted to tackle the event relation extraction problem from a general point of view regardless of the entity type – event, person, location, etc. –, including supervised, unsupervised, semi-supervised, and distant supervision approaches [12]. Each approach has advantages and limitations: supervised approaches heavily rely on large training datasets; on the other hand, unsupervised approaches fall short in labeling the identified clusters – introducing a barrier for human understanding – and in finding unified evaluation metrics; distant supervision approaches are based on entity alignment between a corpus and and a knowledge base, but demonstrated low accuracy scores due to a bad precision.

Particularly, the Cross-Modal Attention Network [13] achieved state-of-the-art performances by simultaneously learning two tasks: entity recognition and relation classification. The approach involves injecting the token-level information into entity tags, rather than concatenating token and label representations.

In the literature, the extraction of events and their relationships studies mostly a subset of possible relations, namely causality, temporality and coreference [5]. Previous studies demonstrated that models based on the combination of CNN, LSTM and attention mechanism are able to capture causal dependencies, even when the cause and effect are separated by a significant distance within the sentence [14].

Event extraction has been made possible using pretrained language models such as BERT [15], as shown in [16]. SpanBERT [17] – an improved version of BERT that excels in predicting text spans instead of single words – has also been employed for event extraction, resulting in notable performance gains [18].

3. Building a Synthetic Event Relations Dataset with GPT-3

The Event Relation dataset (Section 2.1) is the only available dataset including multiple event relation types. In this section, we describe our efforts for overcoming the two most important limitations of this dataset: its size and the large unbalance between relation types.

Our data augmentation strategy for expanding the dataset is based on the automatic generation of sentences using a prompt-based model. Using the right prompt as input, the model would provide new synthetic sentences for enriching the dataset.

We use the GPT-3 language model [11], and more precisely the GPT-3.5 *text-davinci-003* variant as described in the OpenAI documentation.[4] We are interested in generating sentences that involve events and relationships between them, particularly those related to prevention, intention, and enabling.

[4] https://platform.openai.com/docs/guides/completion

3.1. Starting Point: The Event Relations Dataset

The Event Relations Dataset [6] — later named in this paper the *Original Dataset* – describes some of the FARO event relation types. It represents the first events and events relations dataset that encapsulates different event relations, ranging from temporal to causal, and extending beyond causality to include intention, prevention, enabling, and the explicit negation of causality – that we will not cover in this work, because it would require a separate discussion. The construction of the dataset was done by manually re-annotating two existing datasets, TimeBank [7], and [19], which previously only included temporal and causal relations. The dataset was afterwards extended with more samples for prevention and enabling relation using the same manual validation technique.

The annotation of the aforementioned new event relations types involved also the annotation of the constructs of each relation, which we refer to them in the following as *event triggers*. It is worth mentioning that these *event triggers* belong to a bigger class in FARO called *Relata*, which is an abstraction encompassing two sub classes: *Event* (immanent) and *Condition* (transcendent). In this paper, we consider a subset of events that acts as triggers for preventing, causing or intending to cause other events, and a condition that enables the happening of another event.

Example 1. "The **move** boosts Intelogic Chairman Asher Edelman's stake to 20% from 16.2% and may help prevent Martin Ackerman from making a **run** at the computer-services concern."

$$\textit{move} \xrightarrow{\textbf{prevents}} \textit{run}$$

In Example 1., there exists an event relationship of type prevention, and the two event triggers that participate in the relation are **move** and **run** of type *Event.*

Example 2. "The government of Prime Minister Brian Mulroney has been under **pressure** to **reduce** the deficit, which is expected to reach C$30 billion this year."

$$\textit{pressure} \xrightarrow{\textbf{enables}} \textit{reduce}$$

In Example 2., the event relationship involves the type of enabling, with the two corresponding event triggers being **pressure** and **reduce**, in which **pressure** is an event trigger of type *Condition* and **reduce** is of type *Event.*

Table 1 summarizes the number of event relations per relation type in the *Original Dataset* after extending it with news agency samples.

Table 1
Total number of relations in the *Original Dataset.*

Relation type	Cause	Intend	Prevent	Enable	Not-Cause
Number of relations	283	44	89	124	3

3.2. Prompt-based Sample Generation of Sentences

When designing the prompt utilized to generate synthetic examples for a specific relation type, we include:

1. the definition that the FARO ontology assigns to that relation type;
2. a subset of relevant examples from the dataset.

We consider a sequence of words Xi = [x_1, x_{t1}, ..., $_{t2}$, x_n], representing an event relationship occurring between two Relata, of a specific relation type ER_x. The words x_{t1} and x_{t2} respectively represents in the text the two Relata which are the subject and the object of the relations. The definition of the relation type *definition(ER_x)* is taken from the FARO ontology.

The selection of the prompt is done after a series of attempts. For sentences generation, we started by leveraging only the task description in the prompt. Therefore, the generated sentences where too short and basic, while we need realistic and longer sentences, similarly to those in the *Original Dataset*.

Table 2 demonstrates an effort to prompt the model to produce sentences that showcase connection between events with the desired relation type, but the resulting answer falls short of meeting our intended expectations.

The prompt text to generate sample sentences including relations of type ERx is written as the following:

$$\textbf{Prompt(ERx)} = \textit{definition(Event)} + \textit{definition}(\textit{ER}_x) + \textit{request(ER)} + \textit{examples(ERx)}$$

This prompt definition concerns prevention and intention relations. In the context of *enabling* relation, we include the definition of a condition as follows:

$$\textbf{Prompt}(\textbf{ER}_{\textbf{enable}}) = \textit{definition(Event)} + \textit{definition(Condition)} + \textit{definition}(\textit{ER}_{enable}) + \textit{request(ER)} + \textit{examples(ERx)}$$

where *request(ER)* refers to the task description that is given to the language model along with the definitions and *examples(ERx)* are randomly-selected examples from the existing dataset which will be used to iteratively expand and reformulate the dataset.

Example: Prompt used to generate sentences with event relation of type *Enabling*

Note that the original dataset was re-annotated based on Timebank [20] and Event Causality dataset [19], both of which are derived from news articles. This makes the majority of the sentences falling within the political domain. Therefore, introducing the word political in the prompt is to ensure that the generated sentences were coherent and consistent with the original dataset domain.

3.3. Prompt-based Event Trigger Annotation

Similarly to sentence generation, we leverage definitions of events to the prompt, adding few examples illustrating the right position for event triggers for each relation types. The prompts have been chosen to acquire the most similar sample pattern to facilitate parsing.

For an event relationship ERx including prevention or intention, the prompt for selecting their event trigger words is designed as follow:

Table 2
Example of prompting attempts that fell short of producing the desired results

Prompt	**Answer**	**Limitation**
Give me an event that enables the happening of other event.	One event that enables the happening of another event is a person's decision to take an action. For example, a person's decision to get up and walk across a room enables the person to arrive at their destination.	• Prompt had no examples from the existing dataset. • Answer was too simple. • Did not describe a real-world scenario. • Explanation was given instead of an actual expression.
Describe a situation where an event is an intention to cause another event, for example: Companies such as Microsoft or a combined worldcom MCI are trying to monopolize Internet access.	An example of an event intended to cause another event is when companies such as Microsoft or a combined worldcom MCI attempt to monopolize Internet access. By controlling the majority of the market, these companies can dictate the terms and prices of access, potentially limiting consumer choice and driving up costs...	• Explanation was given instead of an actual expression.

definition(Event)	An event is a possible or actual event, which can possibly be defined by precise time and space coordinates.
definition(Condition)	A condition is the fact of having certain qualities, which may trigger events.
definition(ER_{enable})	The *enables relationship* connects a condition or an event (trigger1), with an other event (trigger 2) it is contributing to realize as an enabling factor.
request(ER)	Give me very long political example sentences following these examples and give me each sentence in one line.
examples(ER_{enable})	...

Prompt Event Triggers ERx= definition(Event) + definition(ER_x) + request_{trig}(ERx, sentence, x_{t1}, x_{t2})

where the last element is the description of the task of retrieving event triggers from the text. This request takes the following shape:

If in this sentence <TEXT OF THE SENTENCE> is present an expression with

> a <RELATION TYPE> relationship between <x_{t1}> (trigger1) and <x_{t1}> (trigger2), what would be the trigger1 and trigger2 in these sentences? Give me only one single word for each trigger an only two triggers per sentence. Put each pair between parentheses in a separate line.

For event relations of type *enable*, the definition of the condition is modified in the following way:

Prompt Event Triggers $\mathbf{ER_{enable}}$= definition(Event) + definition(Condition) +
definition($\mathrm{ER_{enable}}$) + $\mathrm{request_{trig}}$($\mathrm{ER_{enable}}$, sentence, $\mathrm{x_{t1}}$, $\mathrm{x_{t2}}$)

Example: Prompt used to generate event triggers with event relation of type *Prevention*

definition(Event)	An event is a possible or actual event, which can possibly be defined by precise time and space coordinates.
definition(Condition)	A condition is the fact of having certain qualities, which may trigger events.
definition($\mathrm{ER_{enable}}$)	The *prevent relationship* connects an event (trigger1) with the event (trigger 2) for which is the cause of not happening.
request(ET)	If in this sentence *"Subcontractors will be offered a settlement and a swift transition to new management is expected to avert an exodus of skilled workers from Waertsilae Marine's two big shipyards, government officials said."* is present an expression with a *prevention* relationship between *settlement* (trigger1) and *exodus* (trigger2), what would be the trigger1 and trigger2 in these sentences? Give me only one single word for each trigger an only two triggers per sentence, put each pair between parentheses in a separate line.

3.4. Manual Validation

We use these methods and we generate 600 sentences with each of the relations.

To guarantee the accuracy of the generated set of samples and their appropriate event triggers, we manually validate each synthetic sentence, ensuring its adherence to the given definition. Overall, 90.77% of all generated sentences were correctly representing an event relation of the requested type. After removing the wrong samples from the dataset, we proceed checking the correctness of their extracted event triggers for the remaining correct sentences.

The generated events triggers were not consistent in term of their patterns from one generation to another. For this reason, an additional parsing step was needed. For doing that, we identified the different textual patterns, processed and categorized these patterns by removing irrelevant words such as '(trigger 1)', and retaining only the precise word or sequence of words that represent the essential part of the event. We were able to identify roughly 12 different patterns. Some examples are reported in Table 3.

After this processing, we validated the correctness of the two trigger words, measuring an accuracy of 75.15% for trigger-1 and 66.82% for trigger-2. Sentences with wrong triggers were

Table 3
Three of the different textual patterns which GPT-3 was returning in output for the Event Triggers selection.

Pattern Number	**Event Triggers**
0	Entitles, Buy
1	Approval (trigger1), Acquire (trigger2)
2	"Trigger1 (military): success Trigger2 (diplomatic): risks"

Table 4
Percentage of Correct Sentences and Event Trigger Words with GPT-3

Relation types	Intention	Prevention	Enabling	Total
Correct Generated Sentences(%)	93.82	97	81.5	90.77
Correct ET1 (%)	75.13	81.83	68.5	75.15
Correct ET2 (%)	73.47	77	50	66.82
Number of Checked Examples	600	600	600	1800

not eliminated from the dataset, but instead manually fixed. Table 4 shows the detailed accuracy scores for each relation type.

We merged the synthetic data with our original dataset, resulting into a larger and more diverse dataset. We managed to acquire and validate 1507 new sentences – with relative event triggers – making a total 2289 sentences. The statistics of this new dataset – latter in the text named the *Augmented Dataset* – are reported in Table 5.

Table 5
Augmented Dataset Statistics

Relation Type	Original Dataset	Augmented Dataset
Prevent	93	646
Enable	118	573
Intend	44	615
Cause	283	283
No-Relation	72	172
total	610	2289

Based on the above reported results, it is evident that GPT-3, with a clear definition of concepts, particularly definition of relations and their constructs types, along with a limited number of examples (5 for each iteration), is able to reach a considerable accuracy. For event triggers generation, we just included a single sentence example, along with its event triggers. Observing the first generated annotations and realising that, despite the limited amount of training data, the results obtained were reasonably good, we decided to continue without adding further sentences. We considered the accuracy quite high, considering the difficulty of the task.

4. Events and Event Relation Extraction

In order to fulfil our goal of testing the effectiveness of GPT-3 based data augmentation technique on events and events relationships extraction, and at the same time, to evaluate the performances of existing models, we conducted the following experiment. We fine-tune two instances of the BERT model [15]. The first one, named BERTee, is for token classification which we apply on event extraction. The second one, named BERTer, is for sequence classification which we apply on events relation classification. We additionally fine-tune a variant of BERT for event extraction, SpanBERT [17] taking into consideration that some of our event triggers are represented by more than one word, i.e. a span of words, and SpanBERT is specifically designed to handle this type of representation.

Xi = [x_1, x_{t1}, ...,x_{t2}, x_n] is a sentence of n tokens, which is part of the studied dataset.

BERTee (Figure 1a) is trained for predicting a tag for each token in the input sequence. The tags are chosen among TAGi=[O,$x1_{type}$,...,$x2_{type,O}$], where $x1_{type}$=[Trigger1] and $x2_{type}$=[Trigger2] are the subject and the object of the event relation, and 'O' is refereed to the rest of tokens in each sentence. The models consists of 12 transformer blocks receiving in input the sequence of tokens Xi and returning the relative contextualized representation H_i= [h1, h2,..., hn]. On top of the transformers, a classification layer – consisting of a fully connected layer followed by a softmax activation – maps each contextualized representation h_i to a probability distribution over the possible labels for that token, i.e., P_{hi} = [P(Trigger1|hi), P(Trigger2|hi), P(O|hi)]. Finally, for x_i ∈Xi, we select the most probable tag TAG_{xi} = $max(P_{hi})$.

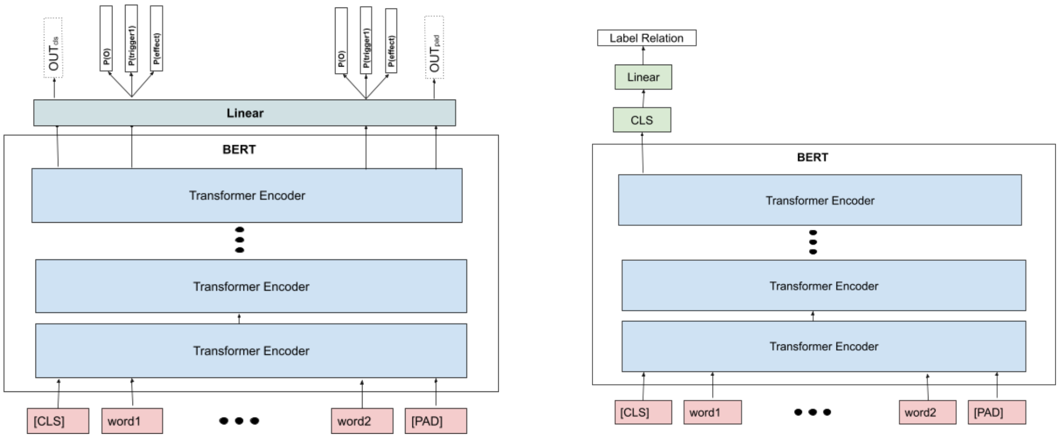
Similarly, **BERTer** (Figure 1b) takes as input the same sequence of tokens Xi, and uses transformers to compute the contextualized representation H_i. This representation feeds a softmax classification layer that maps the hidden states to the event relation types outputting the probability distribution for the label Li ∈L, given L =[causality, enabling, prevention, intention, No-Relation]. We select similarly the most probable label from the outputted labels probabilities.

Some of the event mentions present in our work are denoted as a *span of words*, i.e a sequence of words. Although they present the minority, we wanted to test a variant of BERT called **SpanBERT** [17], which is trained mainly to predict a sequence of words rather than a single word. During the training, the model is masking spans of words rather than random single words, pushing the neural network to predict them. The model is similar to the overall architecture of the BERTee in term of transformers blocks: base model with 12 transformer block. In addition to the classification layer, SpanBERT also includes a span classification layer that is designed to predict the label for a span of tokens. For example, given the sentence: "The United Nations passed a resolution to impose an **arms embargo** on Syria in an effort to pressure the government to end its civil war". The span "**arms embargo**" is aimed to be tagged by SpanBERT as a single class label, namely as "**Trigger1 Trigger1**" more robustly.

5. Results

5.1. Events and Events Relationship Extraction

We trained BERTee, BERTer, and SpanBERT models on both the *Original* and *Augmented* datasets. However, it is important to note that the test set used for evaluation was extracted



(a) BERTee for Event Extraction

(b) BERTer for Event Relation Classification

Figure 1: BERTee and BERTer architectures

solely from the Original dataset. In other words, synthetic data are used only in the training phase, so that they are not distorting the evaluation of the developed systems.

The outcomes of the experiment with BERTer are shown in Table 6. We observe that the performance varied across different classes in the *Original Dataset.* The model showed relatively higher F1 score values for the ‘cause’ and ‘enable’ classes, while it struggled to perform well on ‘intend’ and ‘prevent’ classes. This is probably due to the low support of these latter classes, with intention relation being less represented in the dataset with only 44 example sentences.

Table 6
Event relationship extraction results on the test set with BERTer on both the original and the Augmented dataset

Dataset	Class	Precision	Recall	F1-Score
Original Dataset	cause	0.64	0.83	0.72
	enable	0.77	0.67	0.71
	intend	0.67	0.33	0.44
	prevent	0.62	0.67	0.64
Augmented Dataset	cause	0.75	0.75	0.75 **(+4.16%)**
	enable	0.93	0.96	0.95 **(+33.80%)**
	intend	0.98	0.92	0.95 **(+115.91%)**
	prevent	0.96	0.93	0.94 **(+46.86%)**

Data augmentation leads to significant improvements in the metrics, with an increment of the F1-score for all classes. The highest improvement is observed in the ’prevent’ and ’enable’ classes, and in particular for the ’intend’ class, which received a 115.91% increase in the F1-score with respect to the 44% of the Original Dataset. Despite the absence of added synthetic data for the “cause” relation, we still notice an improvement with a modest margin. Future work will involve data augmentation also for that class.

The performance of BERTee and SpanBERT in events classification is less good in both the Original and Augmented datasets (Table 7). Despite testing different parameters, the results indicated that further improvements are still needed to enhance the performance of this task. Therefore, the current outcomes should be viewed as preliminary, and further investigations will be adopted as future work.

Table 7
Event extraction results on the test set after finetuning BERTee and SpanBERT for token Classification on both the original Dataset and the augmented dataset. (*): The reported metrics are the macro average of Precision, Recall, and F1-score. The macro average F1 score computes the F1 score independently for each class and then takes the unweighted average of the scores. This means that each class is given equal weight, regardless of the number of samples it contains. For this reason, the F1 score values may seem to not reflect the overall precision and recall metrics

Model	Dataset	Precision(*)	Recall(*)	F1-score(*)
SpanBERT	Original Dataset	0.34	0.36	0.12
	Augmented Dataset	0.34	0.32	0.18
BERTee	Original Dataset	0.33	0.31	0.12
	Augmented Dataset	0.33	0.34	0.23

5.2. Modeling the Extracted Event Relations in a Knowledge Graph

Event Knowledge Graphs are shown to be an effective data representation way to ease navigation through event flows and their relations and to flexibly retrieve information about these events from the stored knowledge [21]. This can serve many applications such as link prediction and fact-checking, and their efficiency tends to be considerable when they are richer in terms of aspects and relationships between them. For this sake, we aim to generate a Knowledge Graph (KG) of events and relations between them from the Augmented dataset. In other words, this KG will be an RDF version of the Augmented dataset.

The KG that we constructed contains events and relations between them. The elements in our KG are classified according to the FARO ontology, which distinguishes between two major types of Relata: Condition and Event (see Section 3.1).

More precisely, the events are typed according to the relation between them. 'Enables' relation, was used to connect two entities in which the subject represents the (Condition) that is necessary for the object (Event) to occur. We also identified the other relations that we focus on in the previous parts between events in our KG, such as “causes”, “prevents”, and “intends”, which relate two entities of type Event. These relations capture different types of causal and temporal dependencies between events, and they allow to reason about complex chains of events and their potential consequences.

To ensure the traceability of our KG, we linked each event in our KG to the sentence it was extracted from and each sentence to its provenance corpus, using the Provenance ontology (PROV-O) [22]. This provenance information allows to track the origin of each piece of information in our KG, and to verify its accuracy and relevance. When the provenance involves scientific datasets or software, it is referenced in the graph using the FaBiO ontology [23], detailing information about paper *title*, *author* and *year of publication.*

Additionally, in order to have a more complete KG in terms of event relations (temporal relations, comparative relations, etc.), we have incorporated in the same aforementioned way events and their temporal relations from TimeBank [7] corpus, We also leverage events which have temporal, comparative and contingent relations from the [24] dataset, we call it in the rest of the paper Hong dataset.

TimeBank consists of 24k events with 3.4k temporal links, extracted from 183 News article. On the other hand, Hong consists of the annotation of ACE2005 news-wire documents and other news documents about Malaysian Airline 17, resulting in 862 events with 25610 relations between them.

The integration of the earlier stated datasets was made after examining the overlap between some of their event relation definitions with FARO ontology. The mapping of this relations to FARO is shown in 8 and 9.

The resulting graph – containing over 68,000 statements, with 11,917 event relation links – has been loaded in a triplestore, available for query at http://kflow.eurecom.fr/.

Table 8
Mapping of Event Relation Dataset (Hong) relations to FARO

	FARO		**Hong**	
Category	Super-type	Type	Type	Super-type
Temporal	Temporally related to	Immediately before	Before	Before
		Meets	Before	Meet
		Starts	Starts	Starts
		Ends	Finish	Finish
		Contains(the subproperty)	During	During
		Overlaps	Overlaps	Overlap
		Simulations to	Equality	Equality
Causality	Contingently related to	Causes	Causality	Contingency
Enabling		Enables	Condition	
Comparison	Comparatively related to	Opposite to	Opposite	Comparison
		Alternative to	Negation	
		Contrasting version of	Competition	

6. Conclusion and Future Work

In this work, we made a first attempt towards the automatic extraction of event relations with precise semantics from raw text, focusing on a subset of them. We applied GPT-3 to generate synthetic data for the aforementioned event relation types, obtaining good accuracy. The result of this effort is a dataset consisting of 2289 sentences – of which 1507 were synthetic – annotated with the event mentions in the text. The data augmentation method described in this paper can be used to extract even more event relations by properly replacing the definition and the examples to match the required relation types.

Furthermore, we used BERT for performing two related tasks: event relation classification and event mentions classification. We utilized also SpanBERT to evaluate its ability to classify events that are expressed as a sequence of words into a single class, even though such events are

Table 9
Mapping of TimeBank relations to FARO

Types	**FARO**	**TimeBank**
Temporal	Simultanious to	SIMULTANIOUS
	Before	BEFORE
	Imediately before	IBEFORE
	Inverse of (Imediately before)	IAFTER
	Contains(the subproperty)	During
	Contains	INCLUDES
	Inverse of (contains)	IS INCLUDED
	Starts	BEGINS
	Inverse of (starts)	BEGUN BY
	Ends	ENDS
	Inverse of (ends)	ENDED BY

less frequent. The reported results show that the augmented dataset – and in general synthetic data – improve the ability of the model to generalize and correctly classify sequences, even for classes with a limited number of training examples in the original dataset. However, for event mention classification the performance is still relatively low and needs further improvements.

All code used for the experiments reported in this paper, as well as the resulting dataset is available at https://github.com/ANR-kFLOW/event-relation-classification.

With the recent release of GPT-4 [25], new experiments can be performed for improving the synthetic data generation, in particular in the extraction of relevant triggers.

Furthermore, we would like to investigate the interaction between event relation types and the event mentions by jointly extract them from text in a sense of enhancing the event trigger classification by leveraging event relation information. In particular, We plan to test the effectiveness of the [13]model on our own dataset and assess its performance under similar conditions.

In future work, we intend to combine the automatic classification of event relation to classic event identification techniques, in order to automatically annotate news and encyclopedic entries, with the final goal of realizing a KG of interconnected events with precise semantics.

Acknowledgments

This work has been partially supported by the French National Research Agency (ANR) within the kFLOW project (Grant n°ANR-21-CE23-0028).

References

[1] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals, in: Proceedings of the 5th International Workshop

on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 33–38. URL: https://aclanthology.org/S10-1006.

[2] Z. Nasar, S. W. Jaffry, M. K. Malik, Named Entity Recognition and Relation Extraction: State-of-the-Art, ACM Comput. Surv. 54 (2021). URL: https://doi.org/10.1145/3445965. doi:10.1145/3445965.

[3] D.-T. Vo, E. Bagheri, Extracting Temporal Event Relations Based on Event Networks, in: L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, D. Hiemstra (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2019, pp. 844–851.

[4] V. Khetan, R. Ramnani, M. Anand, S. Sengupta, A. E. Fano, Causal bert: Language models for causality detection between events expressed in text, in: K. Arai (Ed.), Intelligent Computing, Springer International Publishing, Cham, 2022, pp. 965–980.

[5] K. Liu, Y. Chen, J. Liu, X. Zuo, J. Zhao, Extracting events and their relations from texts: A survey on recent research progress and challenges, AI Open 1 (2020) 22–39. doi:https://doi.org/10.1016/j.aiopen.2021.02.004.

[6] Y. Rebboud, P. Lisena, R. Troncy, Beyond Causality: Representing Event Relations in Knowledge Graphs, in: Knowledge Engineering and Knowledge Management (EKAW), Springer International Publishing, Bolzano, Italy, 2022, pp. 121–135. doi:10.1007/978-3-031-17105-5_9.

[7] N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen, J. Pustejovsky, 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), Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 1–9. URL: https://aclanthology.org/S13-2001.

[8] P. Mirza, R. Sprugnoli, S. Tonelli, M. Speranza, Annotating causality in the TempEval-3 corpus, in: Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL), Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 10–19. URL: https://aclanthology.org/W14-0702. doi:10.3115/v1/W14-0702.

[9] X. Feng, J. Guo, B. Qin, T. Liu, Y. Liu, Effective deep memory networks for distant supervised relation extraction, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 4002–4008. doi:10.24963/ijcai.2017/559.

[10] J. Liu, Y. Chen, K. Liu, J. Zhao, Event Detection via Gated Multilingual Attention Mechanism, in: AAAI Conference on Artificial Intelligence, volume 32, 2018, pp. 4865–4872.

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[12] H. Wang, K. Qin, R. Y. Zakari, G. Lu, J. Yin, Deep Neural Network Based Relation Extraction: An Overview, Neural Computing and Applications 34 (2022) 4781–4801. doi:10.1007/s00521-021-06667-3.

[13] S. Zhao, M. Hu, Z. Cai, F. Liu, Modeling dense cross-modal interactions for joint entity-

relation extraction, in: Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 4032–4038.

[14] J. Yang, S. C. Han, J. Poon, A survey on extraction of causal relations from natural language text, Knowledge and Information Systems 64 (2021) 1161 – 1186.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[16] S. Yang, D. Feng, L. Qiao, Z. Kan, D. Li, Exploring pre-trained language models for event extraction and generation, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5284–5294. URL: https://aclanthology.org/P19-1522. doi:10.18653/v1/P19-1522.

[17] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving pre-training by representing and predicting spans, Transactions of the Association for Computational Linguistics 8 (2020) 64–77. URL: https://aclanthology.org/2020.tacl-1.5. doi:10.1162/tacl_a_00300.

[18] B. Portelli, D. Passabi, E. Lenzi, G. Serra, E. Santus, E. Chersoni, Improving adverse drug event extraction with spanbert on different text typologies, ArXiv abs/2105.08882 (2021).

[19] Q. Ning, Z. Feng, H. Wu, D. Roth, Joint Reasoning for Temporal and Causal Relations, in: 56^{th} Annual Meeting of the Association for Computational Linguistics, volume 1, Association for Computational Linguistics, Melbourne, Australia, 2018.

[20] N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen, J. Pustejovsky, SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations, in: 7^{th} International Workshop on Semantic Evaluation (SemEval), Association for Computational Linguistics, Atlanta, USA, 2013, pp. 1–9.

[21] S. Guan, X. Cheng, L. Bai, F. Zhang, Z. Li, Y. Zeng, X. Jin, J. Guo, What is event knowledge graph: A survey, IEEE Transactions on Knowledge & Data Engineering (5555) 1–20. doi:10.1109/TKDE.2022.3180362.

[22] K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, PROV-O: The PROV Ontology, Technical Report, W3C, 2012. URL: http://www.w3.org/TR/prov-o/.

[23] S. Peroni, D. Shotton, FaBiO and CiTO: Ontologies for describing bibliographic resources and citations, Journal of Web Semantics 17 (2012) 33–43. doi:https://doi.org/10.1016/j.websem.2012.08.001.

[24] Y. Hong, T. Zhang, T. O'Gorman, S. Horowit-Hendler, H. Ji, M. Palmer, Building a Cross-document Event-Event Relation Corpus, in: 10th Linguistic Annotation Workshop 2016 (LAW-X 2016), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1–6. doi:10.18653/v1/W16-1701.

[25] OpenAI, GPT-4 Technical Report, 2023. URL: https://arxiv.org/abs/2303.08774. doi:10.48550/ARXIV.2303.08774.