# CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures

**Nasrin Mostafazadeh**$^1$**, Alyson Grealish**$^1$**, Nathanael Chambers**$^2$**,**
**James Allen**$^{1,4}$**, Lucy Vanderwende**$^3$
1 University of Rochester, 2 United States Naval Academy,
3 Microsoft Research, 4 The Institute for Human & Machine Cognition
{nasrinm,agrealis,james}@cs.rochester.edu
nchamber@usna.edu, lucyv@microsoft.com

## Abstract

Learning commonsense causal and temporal relation between events is one of the major steps towards deeper language understanding. This is even more crucial for understanding stories and script learning. A prerequisite for learning scripts is a semantic framework which enables capturing rich event structures. In this paper we introduce a novel semantic annotation framework, called Causal and Temporal Relation Scheme (CaTeRS), which is unique in simultaneously capturing a comprehensive set of temporal and causal relations between events. By annotating a total of 1,600 sentences in the context of 320 five-sentence short stories sampled from ROCStories corpus, we demonstrate that these stories are indeed full of causal and temporal relations. Furthermore, we show that the CaTeRS annotation scheme enables high inter-annotator agreement for broad-coverage event entity annotation and moderate agreement on semantic link annotation.

## 1 Introduction

Understanding events and their relations in natural language has become increasingly important for various NLP tasks. Most notably, story understanding (Charniak, 1972; Winograd, 1972; Turner, 1994; Schubert and Hwang, 2000) which is an extremely challenging task in natural language understanding, is highly dependent on understanding events and their relations. Recently, we have witnessed a renewed interest in story and narrative understanding based on the progress made in core NLP tasks.

Perhaps the biggest challenge of story understanding (and story generation) is having commonsense knowledge for the interpretation of narrative events. This commonsense knowledge can be best represented as scripts. Scripts present structured knowledge about stereotypical event sequences together with their participants. A well known script is the Restaurant Script, which includes the events {Entering, Sitting down, Asking for menus, Choosing meals, etc.}, and the participants {Customer, Waiter, Chef, Tables, etc.}. A large body of work in story understanding has focused on learning scripts (Schank and Abelson, 1977). Given that developing hand-built scripts is extremely time-consuming, there is a serious need for automatically induced scripts (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Balasubramanian et al., 2013; Cheung et al., 2013; Nguyen et al., 2015). It is evident that various NLU applications (text summarization, co-reference resolution and question answering, among others) can benefit from the rich inferential capabilities that structured knowledge about events can provide.

The first step for any script learner is to decide on a corpus to drive the learning process. The most recent resource for this purpose is a corpus of short commonsense stories, called ROCStories (Mostafazadeh et al., 2016), which is a corpus of 40,000 short commonsense everyday stories [1]. This corpus contains high quality[2] five-sentence stories

[1] These stories can be found here: http://cs.rochester.edu/nlp/rocstories

[2] Each of these stories have the following major characteristics: is realistic, has a clear beginning and ending where something happens in between, does not include anything irrelevant

that are full of stereotypical causal and temporal relations between events, making them a perfect resource for learning narrative schemas.

One of the prerequisites for learning scripts from these stories is to extract events and find inter-event semantic relations. Earlier work (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009; Pichotta and Mooney, 2014; Rudinger et al., 2015) defines verbs as events and uses TimeML-based (Pustejovsky et al., 2003) learning for temporal ordering of events. This clearly has many shortcomings, including, but not limited to (1) not capturing a wide range of non-verbal events such as ‘earthquake’, (2) not capturing a more comprehensive set of semantic relations between events such as causality, which is a core relation in stories.

In this paper we formally define a new comprehensive semantic framework for capturing stereotypical event-event temporal and causal relations in commonsense stories, the details of which can be found in Sections 2-4. Using this semantic framework we annotated 320 stories sampled from ROCStories to extract inter-event semantic structures. Our inter-annotator agreement analysis, presented in Section 5 shows that this framework enables high event entity annotation agreement and promising inter-event relation annotation agreement. We believe that our semantic framework better suits the goals of the task of script learning and story understanding, which can potentially enable learning richer and more accurate scripts. Although this work focuses on stories, the CaTeRS annotation framework for capturing inter-event relations can be applied to other genres.

## 2 Event Entities

Our semantic framework captures the set of event entities and their pairwise semantic relations, which together form an inter-connected network of events. In this Section we define event entities and discuss their annotation process.

### 2.1 Definition

Event is mainly used as a term referring to any situation that can happen, occur, or hold. The definition and detection of events has been a topic of interest in various NLP applications. However, there is still no consensus regarding the span of events and how they should be annotated. There has been some good progress in domain-specific annotation of events, e.g., recent Clinical TempEval task (Bethard, 2013) and THYME annotation scheme (Styler et al., 2014), however, the detection of events in broad-coverage natural language has been an ongoing endeavor in the field.

One of the existing definitions for event is provided in the TimeML annotation schema (Pustejovsky et al., 2003):

> “An event is any situation (including a process or state) that happens, occurs, or holds to be true or false during some time point (punctual) or time interval (durative).”

According to this definition, adjectival states such as ‘on board’ are also annotated as events. As we are focused on the task of narrative structure learning, we want to capture anything that ‘happens and occurs’, but not including holds. Formally, we therefore define an event as follows:

> “An event is any situation (including a process or state) that happens or occurs either instantaneously (punctual) or during a period of time (durative).”

In order to make the event annotation less subjective, we specifically define an event to be any lexical entry under any of the following ontology types in the TRIPS ontology[3] (Allen et al., 2008):

– Event-of-state: e.g., have, lack.

– Event-of-change: e.g., kill, delay, eat.

– Event-type: e.g., effort, procedure, turmoil, mess, fire.

– Physical-condition: all medical disorders and conditions, e.g., cancer, heart attack, stroke, etc.

– Occurring: e.g., happen, occur.

– Natural-phenomenon: e.g., earthquake, tsunami.

This ontology has one of the richest event hierarchies, which perfectly serves our purpose of broad-coverage event extraction.

### 2.2 How to annotate events?

After pinpointing an event entity according to the formal definition presented earlier, one should annotate the event by selecting the corresponding span

to the core story.

[3] http://www.cs.rochester.edu/research/trips/lexicon/browse-ont-lex-ajax.html

in the sentence. Here, we define the event span to be the head of the main phrase which includes the complete event. For example, in the sentence 'I [climbed] the tree', we annotate 'climb', the head of the verb phrase, while the complete event is '[climb] tree' . This implies that only main events (and not their dependents) are annotated. Annotating the head word enables us to delegate the decision about adding the dependent children to a post-process, which can be tuned per application.

Moreover, no verbs which take the role of an auxiliary verb are annotated as events. For instance, in the sentence 'She had to [change] her jeans' the main event is 'change'. For multi-word verbs such as 'turn out or 'find out', the entire span should be selected as the event. The annotators can consult lexicons such as WordNet (Miller., 1995) for distinguishing multi-word verbs from verbs with prepositional phrases adjuncts.

### 2.2.1 The Case for Embedded Events

Another important controversial issue is what to do with embedded events, where one event takes another event as its core argument (if neither of the verbs are auxiliary). For instance, consider the following example:

(1) Sam [wanted]$_{e1}$ to [avoid]$_{e2}$ any trouble, so he [drove]$_{e3}$ slowly.

According to our event entity definition, there are three events in example 1, $e1$ and $e2$ and $e3$, all of which should be annotated. However, more complicated is the case of a main event in an embedded event construction which signals any of the semantic relations in the annotation scheme. Consider the sentence in example 2. In this sentence, there are also three events according to our definition of event entities, where (cause (die)) is an embedded event construction. In this case the verb 'cause' simply signals a causal relation between $e1$ and $e3$, which will be captured by our existing semantic relation (to be described in Section 3), and so we do not annotate the verb 'cause' as an event.

Likewise, the sentence in example 3 showcases another embedded event construction, (cause (explosion)), so the event 'cause' should not be annotated.

(2) The [explosion]$_{e1}$ caused him to [die]$_{e3}$.

(3) The [fire]$_{e1}$ caused an [explosion]$_{e2}$.

The same rule applies to the synonyms of these verbs in addition to other verbs that signal a temporal or causal relation, including but not limited to {start, begin, end, prevent, stop, trigger}, which hereinafter we call 'aspectual verbs'[4]. It is important to note that the above rule for aspectual verbs can be applied only to embedded event constructions and may be overridden. Consider example 4. In this example, it is clear that the event 'prevent' plays a key semantic role in the sentence and should be annotated as an event since it is the only viable event that can be semantically connected to other events such as $e3$.

(4) John [prevented]$_{e1}$ the vase from [falling]$_{e2}$ off the table, I was [relieved]$_{e3}$.

## 2.3 The Case for Copulas

A copula is a verb which links the subject of a sentence with a predicate, such as the verb 'is' which links 'this suit' to the predicate 'dark blue' in the sentence 'This suit is dark blue'. Many such constructions assign a state to something or someone which holds true for some duration. The question is what to specify as the event entity in such sentences. According to our definitions, an adjective such as 'blue' is not an event (that is, it does not occur or happen), but after many rounds of pilot annotations, we concluded that annotating the predicate adjective or predicate nominal best captures the core semantic information. Thus, the sentences 5-6 will be annotated as follows:

(5) He was really [hungry]$_{e1}$.

(6) He [ate]$_{e1}$ a juicy burger.

It is important to emphasize that annotating states such as 'hungry' as an event is only done in the case of copulas, and, for example in sentence 6, the adjective 'juicy' will not be annotated as an event.

Our annotation of light verb constructions (e.g., do, make, have) is consistent with the annotation of copulas and auxiliary verbs. Whenever the semantic contribution of the verb is minimal and the non-verb element of the construction is an event in the

[4] These verbs are the same as aspectual events characterized by TimeML, which include 'INITIATES', 'CULMINATES', 'TERMINATES', 'CONTINUES' and 'REINITIATES'.

TRIPS ontology, we annotate the non-verb element as the event. Thus, we annotate the noun predicate 'offer' in the sentence 'Yesterday, John made an offer to buy the house for 350,000', similarly to the way Abstract Meaning Representation (AMR) drops the light verb and promotes the noun predicate (Banarescu et al., 2013). This annotation is also close to the PropBank annotation of copulas and light verbs (Bonial et al., 2014), where they annotate the noun predicate and predicate adjective as the event; however, PropBank includes an explicit marking of the verb as either a light verb or a copula verb.

## 3 The Semantic Relations Between Event Entities

A more challenging problem than event entity detection is the identification of the semantic relation that holds between events. Events take place in time, hence temporal relations between events are crucial to study. Furthermore, causality plays a crucial role in establishing semantic relation between events, specifically in stories. In this Section, we provide details on both temporal and causal semantic relations.

### 3.1 Temporal Relation

Time is the main notion for anchoring the changes of the world triggered by sequences of events. Of course having temporal understanding and temporal reasoning capabilities is crucial for many NLP applications such as question answering, text summarization and many others. Throughout the years the issue of temporal analysis and reasoning in natural language has been addressed via different approaches. Allen's Interval Algebra (Allen, 1984) is one theory for representing actions and introduces a set of 13 distinct, exhaustive, and qualitative temporal relations that can hold between two time intervals. The first three columns of Table 1 list these 13 temporal relations together with their visualization, which includes 6 main relations and their corresponding inverses –together with the 'equal' relation which does not have an inverse.

Based on Interval Algebra, a new markup language for annotating events and temporal expressions in natural language was proposed (Pustejovsky et al., 2003), named TimeML. This schema is designed to address problems in event and temporal expression markup. It covers two major tags: 'EVENT' and 'TIMEX3'. The EVENT tag is used to annotate elements in a text that represent events such as 'kill' and 'crash'. TIMEX is mainly used to annotate explicit temporal expressions, such as times, dates and durations. One of the major features introduced in TimeML was the LINK tag. These tags encode the semantic relations that exist between the temporal elements annotated in a document. The most notable LINK is TLINK: a Temporal Link representing the temporal relationship between entities (events and time expressions). This link not only encodes a relation between two entities, but also makes a partial ordering between events. There are 14 TLINK relations in TimeML, adding 'simultaneous' to the list of temporal links between events. The fourth column of Table 1 shows the correspondence between Allen relations and TimeML TLINKs. Furthermore, in the fifth column we include the THYME annotation schema (to be discussed in Section 6).

We propose a new set of temporal relations for capturing event-event relations. Our final set of temporal relations are shown in the sixth column of Table 1[5]. As compared with TimeML, we drop the relations 'simultaneous', 'begins' and 'ends'. 'Simultaneous' was not a part of the original Allen relations. Generally it is hard to be certain about two events occurring exactly during the same time span, starting together and ending together. Indeed, the majority of events which are presumed 'simultaneous' in TimeML annotated corpora are either (1) EVENT-TIMEX relations which are not event-event relations, or (2) wrongly annotated and should be the 'overlapping' relation, e.g., in the following sentence from TimeBank corpus the correct relation for the two events $e1$ and $e2$ should be 'overlap':

She [listened]$_{e1}$ to music while [driving]$_{e2}$.

We acknowledge that having 'simultaneous' can make the annotation framework more comprehensive and may apply in few certain cases of punctual events, however, such cases are very rare in our corpus, and in the interest of a more compact and less ambiguous annotation, we did not include

[5] Since a main temporal relation and its inverse have a reflexive relation, the annotation is carried out only on the main temporal relation.

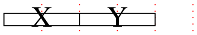
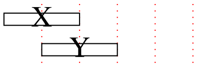
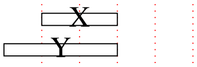
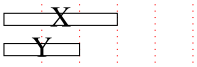
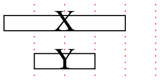
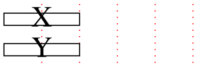
| Allen | Visualization | Allen - Inverse | TimeML | THYME | CaTeRS |
|---|---|---|---|---|---|
| X Before Y | | Y After X | Before | Before | Before |
| X Meets Y |  | Y Is Met X | IBefore (Immediately) | - | - |
| X Overlaps Y |  | Y Is overlapped by X | - | Overlaps | Overlaps |
| X Finishes Y |  | Y Is finished by X | Ends | Ends-on | - |
| X Starts Y |  | Y Is started by X | Begins | Begins-on | - |
| X Contain Y |  | Y During X | During | Contains | Contains |
| X Equals Y | | - | Identity | Identity | Identity |
| - |  | - | Simultaneous | - | - |

Table 1: The correspondence of temporal relation sets of different annotation frameworks.

it. THYME also dropped 'simultaneous' for similar reasons.

As for the 'begins' and 'ends', our multiple pilot studies, indicated that these relations are more accurately captured by one of our causal relations (next subsection) or the relation 'overlaps'. We believe that our simplified set of 4 temporal relations can be used for any future broad-coverage inter-event temporal relation annotation. We also drop the temporal relation 'IBefore', given that this relation usually reflects on causal relation between two events which will be captured by our causal links.

### 3.2 Causal Relation

Research on the extraction of event relations has concerned mainly the temporal relation and time-wise ordering of events. A more complex semantic relationship between events is causality. Causality is one of the main semantic relationships between events where an event (CAUSE) results in another event (EFFECT) to happen or hold. It is clear that identifying the causal relation between events is crucial for numerous applications, including story understanding. Predicting occurrence of future events is the major benefit of causal analysis, which can itself help risk analysis and disaster decision making. There is an obvious connection between causal relation and temporal relation: by definition, the CAUSE event starts 'BEFORE' the EFFECT event. Hence, predicting causality between two events also requires/results in a temporal prediction.

It is challenging to define causality in natural language. Causation, as commonly understood as a notion for understanding the world in the philosophy and psychology, is not fully predicated in natural language (Neeleman and Koot, 2012). There have been several attempts in the field of psychology for modeling causality, e.g., the counterfactual model (Lewis, 1973) and the probabilistic contrast model (Cheng and Novick, 1992). Leonard Talmy's seminal work (Talmy, 1988) in the field of cognitive linguistics models the world in terms of semantic categories of how entities interact with respect to force (Force Dynamics). These semantic categories include concepts such as the employment of force, resistance to force and the overcoming of resistance, blockage of a force, removal of blockage, and etc. Force dynamics provides a generalization over the traditional linguistic understanding of causation by categorizing causation into 'letting', 'helping', 'hindering' and etc. Wolff and Song (Wolff and Song, 2003) base their theory of causal verbs on force dynamics. Wolff proposes (Wolff, 2007) that causation includes three main types of *causal concepts*: **'Cause', 'Enable'** and **'Prevent'**. These three causal concepts are lexicalized through distinctive types of verbs (Wolff and Song, 2003) which are as follows:

– Cause-type verbs: e.g. cause, start, prompt, force.

– Enable-type verbs: e.g. allow, permit, enable, help.

– Prevent-type verbs: e.g. block, prevent, hinder, restrain.

Wolff's model accounts for various ways that causal concepts are lexicalized in language and we base our annotation framework on this model. However, we will be looking at causal relation between events more from a **'commonsense reasoning'** perspective than linguistic markers. We define cause, enable and prevent for commonsense co-occurrence of events, inspired by mental model theory of causality (Khemlani et al., 2014), as follows:

– A $Cause$ B: In the context, If A occurs, B most probably occurs as a result.

– A $Enable$ B: In the context, If A does not occur, B most probably does not occur (not enabled to occur).

– A $Prevent$ B: In the context, If A occurs, B most probably does not occur as a result.

where *In the context* refers to the underlying context in which A and B occur, such as a story. This definition is in line with the definition of CAUSE and PRECONDITION presented in the RED annotation guidelines (Ikuta et al., 2014) (to be discussed in Section 6).

In order to better understand the notion of commonsense causality, consider the sentences 7-9.

(7) Harry [fell]$_{e1}$ and [skinned]$_{e2}$ his knee.

(8) Karla [earned]$_{e1}$ more money and finally [bought]$_{e2}$ a house.

(9) It was [raining]$_{e1}$ so hard that it prevented me from [going]$_{e2}$[6] to the school.

In the above three sentences, the relation between $e1$ and $e2$ is 'cause', 'enable' and 'prevent' in order.

It is important to note that our scope of lexical semantic causality only captures events causing events (Davidson, 1967), however, capturing individuals as cause (Croft, 1991) is a another possible extension.

### 3.2.1 Temporal Implications of Causality

The definition of causality implies that when A Causes B, then A should start before B in order to have triggered it. It is important to note that for durative events the temporal implication of causality is mainly about the start of the causal event, which should be before the start of the event which is caused. Consider the following example:

(10) The [fire]$_{e1}$ [burned down]$_{e2}$ the house.

In this example there is a 'cause' relation between $e1$ and $e2$, where 'fire' clearly does not finish before starting of the 'burn' event. So the temporal relation between the two events is 'overlaps'. Here we conclude that when 'A cause/enable/prevent B', we know as a fact that As start is before Bs start, but there is no restriction on their relative ending. This implies that a cause relation can have any of the two temporal relations: *before* and *overlaps*.

All the earlier examples of causal concepts we explored involved an event causing another event to happen or to start (in case of a durative event). However, there are examples of causality which involve not starting but ending an ongoing event. Consider the sentence 11. In order to capture causality relations between pairs of events such as $e1$ and $e2$, we introduce a *Cause-to-end* relation, which can have one of the three temporal implications: *before*, *overlaps*, and *during*. Hence, in sentence 11, the relation between $e1$ and $e2$ will be *Cause-to-end (overlap).*

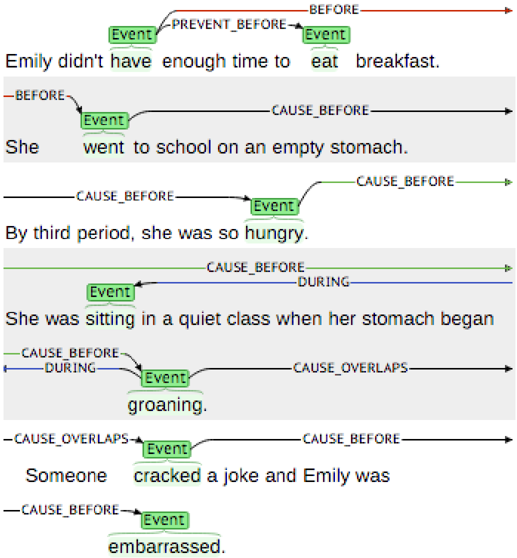
(11) The [famine]$_{e1}$ ended the [war]$_{e2}$.

## 3.3 How to annotate semantic relations between events?

In summary, the disjunctive[7] set of 13 semantic relations between events in our annotation framework are as follows:

[6] As discussed earlier, here the embedded event construction is (prevent (going)) where only the event 'going' will be annotated.

[7] This implies that only one of these semantic relations can be selected per each event pair.



Emily didn't have enough time to eat breakfast.
She went to school on an empty stomach.
By third period, she was so hungry.
She was sitting in a quiet class when her stomach began groaning.
Someone cracked a joke and Emily was embarrassed.

Figure 1: Semantic annotation of a sample story.

- **9 causal relations:** Including 'cause (before/overlaps)', 'enable (before/overlaps)', 'prevent (before/overlaps)', 'cause-to-end (before/overlaps/during)'
- **4 temporal relations:** Including 'Before', 'Overlaps', 'Contains', 'Identity'.

The semantic relation annotation between two events should start with deciding about any causal relations and then, if there was not any causal relation, proceed to choosing any existing temporal relation.

## 4 Annotating at Story level

It has been shown (Bittar et al., 2012) that temporal annotation can be most properly carried out by taking into account the full context for sentences, as opposed to TimeML, which is a surface-based annotation. The scope and goal of this paper very well aligns with this observation. We carry out the annotation at the story level, meaning that we annotate inter-event relations across the five sentences of a story. It suffices to do the event-event relation specification minimally given the transitivity of temporal relations. For example for three consecutive events $e1$ $e2$ $e3$ one should only annotate the 'before' relation between $e1$ and $e2$, and $e2$ and $e3$, since the 'before' relation between $e1$ and $e3$ can be inferred from the other two relations. The event-event relations can be from/to any sentence in the story. It is important to emphasize here that the goal of annotation is to capture commonsensical relations between events, so the annotation effort should be driven by intuitive commonsensical relation one can pinpoint throughout a story.

Consider an example of a fully annotated story shown in Figure 1. As you can see, all of the semantic annotations reflect commonsense relation between events given the underlying story, e.g., 'cracked a joke' *cause-before* 'embarrassed'.

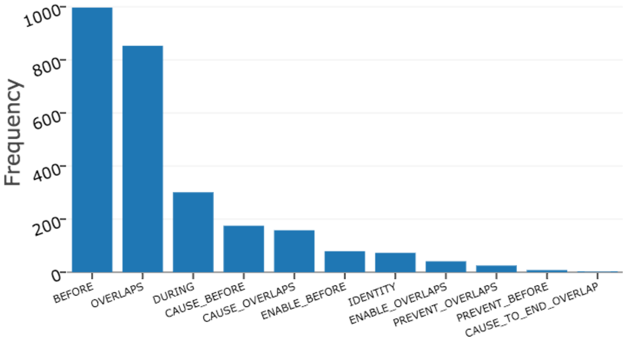


Figure 2: Frequency of semantic links in our dataset.

## 5 Annotated Dataset Analysis

We randomly sampled 320 stories (1,600 sentences) from the ROCStories Corpus. This set covers a variety of everyday stories, with titles ranging from 'got a new phone' to 'Left at the altar'. We provided our main expert annotator with the annotation guidelines, with the task of annotating each of the 320 stories at story level. The annotation task was set up on Brat tool[8]. On average, annotation time per story was 11 minutes. These annotations can be found through `http://cs.rochester.edu/nlp/rocstories/CaTeRS/`.

### 5.1 Statistics

Overall, we have annotated 2,708 event entities and 2,715 semantic relations. Figure 2 depicts the distribution of different semantic relations in our annotation set. For this Figure we have removed any semantic relations with frequency less than 5. As you can see, the temporal relation *before* is the most common relation, which reflects on the natural reporting of the sequence of events throughout a

[8] `http://brat.nlplab.org/`

story. There are overall 488 various causal links, the most frequent of which is *cause-before*. Given these statistics, it is clear that capturing causality along with temporal aspects of stories is crucial.

Our annotations also enable some deeper analysis of the narrative flow of the stories. One major question is if the text order of events appearing in consecutive sentences mirrors the real-world order of events. Although the real-world order of events is more complex than just a sequence of *before* relations, we can simplify our set of semantic links to make an approximation: we count the number of links (from any[9] type) which connect an event entity appearing in position $X$ to an event entity appearing in position $X - i$. We found that temporal order does not match the text order 23% of the time. This reinforces the quality of the narrative flow in the ROCStories corpus. Moreover, 23% is statistically significant enough to motivate the requirement for temporal ordering models for these stories.

## 5.2 Inter-annotator Agreement

In order to compute inter-annotator agreement on our annotation framework, we shared one batch of 20 stories between four expert annotators. Our annotation task consists of two subtasks: (1) entity span selection: choosing non-overlapping event entity spans for each story, (2) semantic structure annotation: building a directed graph (most commonly connected) on top of the entity spans.

### 5.2.1 Agreement on Event Entities

Given that there are no prefixed set of event entity spans for straight-forward computation of inter-annotator agreement, we do the following: among all the annotators, we aggregate the spans of the annotated event entity as the annotation object (Artstein and Poesio, 2008). Then, if there exists a span which is not annotated by one of the coders (annotators) it will be labeled as 'NONE' for its category. The agreement according to Fleiss's Kappa $\kappa = 0.91$, which shows substantial agreement on event entity annotation. Although direct comparison of $\kappa$ values is not possible, as a point of reference, the event span annotation of the most recent clinical TempEval (Bethard et al., 2015) was 0.81.

### 5.2.2 Agreement on Semantic Links

Decisions on semantic links are dependent on two things (1) decisions on event entities; (2) the decision about the other links. Hence, the task of annotating event structures is in general a hard task. In order to relax the dependency on event entities, we fix the set of entities to be the ones that all annotators have agreed on. Following the other discourse structure annotation tasks such as Rhetorical Structure Theory (RST), we aggregate all the relations captured by all annotators as the annotation object, then labeling 'NONE' as the category for coders who have not captured this relation. The agreement according to Fleiss's Kappa $\kappa = 0.49$ without applying basic closure and $\kappa = 0.51$ with closure[10], which shows moderate agreement. For reference, the agreement on semantic link annotation in the most recent clinical TempEval was 0.44 without closure and 0.47 with closure.

# 6 Related Work

One of the most recent temporal annotation schemas is Temporal Histories of Your Medical Event (THYME) (Styler et al., 2014). This annotation guideline was devised for the purpose of establishing timelines in clinical narratives, i.e. the free text portions contained in electronic health records. In their work, they combine the TimeML annotation schema with Allen Interval Algebra, identifying the five temporal relations BEFORE, OVERLAP, BEGINS-ON, ENDS-ON, and CONTAINS. Of note is that they adopt the notion of narrative containers (Pustejovsky and Stubbs, 2011), which are time slices in which events can take place, such as DOCTIME (time of the report) and before DOCTIME. As such, the THYME guideline focuses on ordering events with respect to specific time intervals, while in our work, we are only focused on the relation between two events, without concern for ordering. Their simplification of temporal links is similar to ours, however, our reasoning for simplification takes

[9] This is based on the fact that any relation such as 'A enable-before B' or 'A overlaps B' can be naively approximated to 'A before B'.

[10] Temporal closure (Gerevini et al., 1995) is a reasoning for deriving explicit relations to implicit relations, applying rules such as transitivity.

into account the existence of causality, which is not captured by THYME.

Causality is a notion that has been widely studied in psychology, philosophy, and logic. However, precise modeling and representation of causality in NLP applications is still an open issue. A formal definition of causality in lexical semantics can be found in (Hobbs, 2005). Hobbs introduces the notion of "causal complex", which refers to some collection of eventualities (events or states) for which holding or happening entails the happening of effect. In part, our annotation work is motivated to learn what the causal complexes are for a given event or state. The Penn Discourse Tree Bank (PDTB) corpus (Prasad et al., 2008) addresses the annotation of causal relations, annotating semantic relations that hold between exactly two Abstract Objects (called Arg1 and Arg2), expressed either explicitly via lexical items or implicitly via adjacency in discourse. In this paper, we present a semantic framework that captures both explicit and implicit causality, but no constraint of adjacency is imposed, allowing commonsense causality to be captured at the inter-sentential level (story).

Another line of work annotates temporal and causal relations in parallel (Steven Bethard and Martin, 2008). Bethard et al. annotated a dataset of 1,000 conjoined-event temporal-causal relations, collected from Wall Street Journal corpus. Each event pair was annotated manually with both temporal (BEFORE, AFTER, NO-REL) and causal relations (CAUSE, NO-REL). For example, sentence 12 is an entry in their dataset. This dataset makes no distinction between various types of causal relation.

(12) Fuel tanks had leaked and contaminated the soil.
- (leaked BEFORE contaminated)
- (leaked CAUSED contaminated).

A recent work (Mirza and Tonelli, 2014) has proposed a TimeML-style annotation standard for capturing causal relations between events. They mainly introduce 'CLINK', analogous to 'TLINK' in TimeML, to be added to the existing TimeML link tags. Under this framework, Mirza et al (Mirza and Tonelli, 2014) annotates 318 CLINKs in TempEval-3 TimeBank. They only annotate explicit causal relations signaled by linguistic markers, such as {because of, as a result of, due to, so, therefore, thus}. Another relevant work is Richer Event Descriptions (RED) (Ikuta et al., 2014), which combines event coreference and THYME annotations, and also introduces cause-effect annotation in adjacent sentences to achieve a richer semantic representation of events and their relations. RED also distinguishes between 'PRECONDITION' and 'CAUSE', similarly to our 'ENABLE' and 'CAUSE' relations. These can be in the context of BEFORE or OVERLAP, but they do not include PREVENT and CAUSE-TO-END. Our set of comprehensive 9 causal relations distinguishes between various temporal implications, not covered by any of the related work.

## 7 Conclusion

In this paper we introduced a novel framework for semantic annotation of event-event relations in commonsense stories, called CaTeRS. We annotated 1,600 sentences throughout 320 short stories sampled from ROCStories corpus, capturing 2,708 event entities and 2,715 semantic relations, including 13 various types of causal and temporal relations. This annotation scheme is unique in capturing both temporal and causal inter-event relations. We show that our annotation scheme enables high inter-annotator agreement for event entity annotation. This is due to our clear definition of events which (1) is linked to lexical entries in TRIPS ontology, removing problems caused by each annotator devising his or her own notion of event, (2) captures only the head of the underlying phrase.

Our inter-annotator analysis of semantic link annotation shows moderate agreement, competitive with earlier temporal annotation schemas. We are planning to improve the agreement on semantic links even further by setting up two-stage expert annotation process, where we can pair annotators for resolving disagreements and modifying the annotation guideline. A comprehensive study of temporal and causal closure in our framework is a future work.

We believe that our semantic framework for temporal and causal annotation of stories can better model the event structures required for script and narrative structure learning. Although this work focuses on stories, our annotation framework for capturing inter-event relations can be applicable to other genres. It is important to note that this framework is not intended to be a comprehensive analysis of all temporal and causal aspects of text, but rather we focus on those temporal and causal links that support learning stereotypical narrative structures. As with any annotation framework, this framework will keep evolving over time, the updates of which can be followed through http://cs.rochester.edu/nlp/rocstories/CaTeRS/.

## Acknowledgments

We thank Martha Palmer and the anonymous reviewers for their invaluable and extensive comments on this paper. This work was supported in part by Grant W911NF-15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA), the Army Research Office (ARO) and Office of Naval Research (ONR).

## References

James F. Allen, Mary Swift, and Will de Beaumont. 2008. Deep semantic analysis of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 343–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

James F. Allen. 1984. Towards a general theory of action and time. *Artif. Intell.*, 23(2):123–154, July.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December.

Niranjan Balasubramanian, Stephen Soderland, Oren Etzioni Mausam, and Oren Etzioni. 2013. Generating coherent event schemas at scale. In *EMNLP*, pages 1721–1731.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado, June. Association for Computational Linguistics.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA. Association for Computational Linguistics.

Andr Bittar, Caroline Hagge, Vronique Moriceau, Xavier Tannier, and Charles Teissdre. 2012. Temporal annotation: A proposal for guidelines and an experiment with inter-annotator agreement. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Claire Bonial, Julia Bonn, Kathryn Conger, Jena D. Hwang, and Martha Palmer. 2014. Propbank: Semantics of new predicate types. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Nathanael Chambers and Daniel Jurafsky. 2008. Unsupervised learning of narrative event chains. In Kathleen McKeown, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *ACL*, pages 789–797. The Association for Computer Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. 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: Volume 2 - Volume 2*, ACL '09, pages 602–610, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eugene Charniak. 1972. Toward a model of children's story comprehension. December.

Patricia W. Cheng and Laura R. Novick. 1992. Covariation in natural causal induction. *Psychological Review*, 99(2):365382.

Jackie Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. In *ACL*.

William. Croft. 1991. *Syntactic categories and grammatical relations : the cognitive organization of information / William Croft*. University of Chicago Press Chicago.

Donald Davidson. 1967. Causal relations. *Journal of Philosophy*, 64(21):691–703.

Alfonso Gerevini, Lenhart K. Schubert, and Stephanie Schaeffer. 1995. The temporal reasoning tools timegraph i-ii. *International Journal on Artificial Intelligence Tools*, 4(1-2):281–300.

Jerry R. Hobbs. 2005. Toward a useful concept of causality for lexical semantics. *Journal of Semantics*, 22(2):181–209.

Rei Ikuta, Will Styler, Mariah Hamang, Tim O'Gorman, and Martha Palmer. 2014. Challenges of adding causation to richer event descriptions. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 12–20, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Sangeet Khemlani, Aron K Barbey, and Philip Nicholas Johnson-Laird. 2014. Causal reasoning with mental models. *Frontiers in Human Neuroscience*, 8(849).

David Lewis. 1973. *Counterfactuals*. Blackwell Publishers, Oxford.

G. Miller. 1995. Wordnet: A lexical database for english. In *In Communications of the ACM*.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2097–2106.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL HLT*, San Diego, California, June. Association for Computational Linguistics.

Ad Neeleman and Hans Van De Koot. 2012. The theta system: Argument structure at the interface. *The Linguistic Expression of Causation*, pages 20–51.

Kiem-Hieu Nguyen, Xavier Tannier, Olivier Ferret, and Romaric Besançon. 2015. Generative event schema induction with entity disambiguation. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics (ACL-15)*.

Karl Pichotta and Raymond J Mooney. 2014. Statistical script learning with multi-argument events. *EACL 2014*, page 220.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The penn discourse treebank 2.0. In *LREC*. European Language Resources Association.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, LAW V '11, pages 152–160, Stroudsburg, PA, USA. Association for Computational Linguistics.

James Pustejovsky, Jos Castao, Robert Ingria, Roser Saur, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003. Timeml: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5*.

Rachel Rudinger, Pushpendre Rastogi, Francis Ferraro, and Benjamin Van Durme. 2015. Script induction as language modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP-15)*.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ.

Lenhart K. Schubert and Chung Hee Hwang. 2000. Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. In *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. MIT/AAAI Press.

Sara Klingenstein Steven Bethard, William Corvey and James H. Martin. 2008. Building a corpus of temporal-causal structure. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

IV William F. Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Leonard Talmy. 1988. Force dynamics in language and cognition. *Cognitive Science*, 12(1):49–100.

Scott R. Turner. 1994. The creative process: A computer model of storytelling. *Hillsdale: Lawrence Erlbaum.*

Terry Winograd. 1972. *Understanding Natural Language*. Academic Press, Inc., Orlando, FL, USA.

Phillip Wolff and Grace Song. 2003. Models of causation and the semantics of causal verbs. *Cognitive Psychology*, 47(3):276–332.

Phillip Wolff. 2007. Representing causation. *Journal of Experiment Psychology: General*, 136:82111.