# NarrativeTime: Dense Temporal Annotation on a Timeline

**Anna Rogers[♣], Marzena Karpinska[♡], Ankita Gupta[♡], Vladislav Lialin[♠], Gregory Smelkov[♠], Anna Rumshisky[♠]**
[♣]IT University of Copenhagen, [♡]UMass Amherst, [♠]UMass Lowell
arog@itu.dk, {mkarpinska, ankitagupta}@cs.umass.edu, {vlialin, gsmelkov, arum}@cs.uml.edu

**Abstract**

For the past decade, temporal annotation has been sparse: only a small portion of event pairs in a text was annotated. We present NarrativeTime, the first timeline-based annotation framework that achieves full coverage of all possible tlinks. To compare with the previous SOTA in dense temporal annotation, we perform full re-annotation of the classic TimeBankDense corpus (American English), which shows comparable agreement with a signigicant increase in density. We contribute TimeBankNT corpus (with each text fully annotated by two expert annotators), extensive annotation guidelines, open-source tools for annotation and conversion to TimeML format, and baseline results.

**Keywords:** temporal annotation, TimeBank, event order

## 1. Introduction

Event order information is usually represented by temporal links (tlinks) between events pairs: does $event_1$ happen Before/During/After $event_2$? Ideally, temporal annotation would establish *all* tlinks in the text, but since their number is quadratic to the number of events in the text, it is usually *sparse*: e.g. TimeBank only contains 1-5% of all possible tlinks (Verhagen, 2005). Furthermore, much of this information is underspecified in the text, and is not normally inferred by human readers (nor do they make the same inferences if pressed to do so). Several solutions have been proposed for the density problem (Verhagen, 2005; Cassidy et al., 2014) and for the underspecification problem (Bethard et al., 2012; Ning et al., 2018), but they remain a challenge.

We address both of these problems in NarrativeTime, the first *timeline-based framework for full temporal annotation*. While the traditional TimeBank-style annotation focuses on relations in individual event pairs, partly annotated and partly inferred (Figure 1a), NarrativeTime builds a dynamic timeline (Figure 1b). That representation is equivalent to the full set of all possible tlinks in the text, and they are guaranteed to be backed by manual annotation (which may not be the case for the pairwise approach). Its solutions to the underspecification problem is based on three mechanisms: event types, timeline branches and factuality.

We implement NarrativeTime framework in detailed annotation guidelines and open-source tools[1] for annotation and conversion to the standard TimeML format. For direct comparison between our approach and prior work, we re-annotate TimeBank-Dense (Cassidy et al., 2014) corpus (American English), with each document fully and independently annotated by two expert annotators.

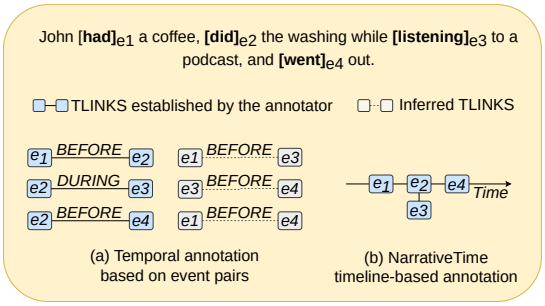



Figure 1: Timeline-based annotation vs annotation based on event pairs.

We achieve inter-annotator agreement (IAA) of Krippendorff's $\alpha$ 0.68 (Krippendorff, 2004). This is comparable or superior to what is reported in the prior work on news texts, but NarrativeTime annotation is dense: it yields 102,313 tlinks[2] vs 12,715 tlinks in the original TimeBankDense (Cassidy et al., 2014) and 1,341 tlinks in the same files in the original TimeBank (Pustejovsky et al., 2003b). We also contribute initial modeling results for temporal relation classification based on LongT5 (Guo et al., 2021) encoder, which suggest that the task is challenging, and there is room for improvement.

To clarify the terminology: we use the term *framework* to differentiate between annotation workflows that are based on relations between individual event pairs, and timeline-based annotation. *Annotation scheme* refers to the specific set of policies about what to annotate and how, which is implemented in *annotation guidelines*. Both timeline- and event-pair-based frameworks can support different

[1]Annotation guidelines, tools, and annotated data are available under MIT license at https://github.com/text-machine-lab/nt

[2]TimeBankNT contains 2 full sets of annotations, each with 102,313 tlinks excluding inverses (symmetrical tlinks that can be auto-inferred, such as X Before Y → Y After X), and 204,626 tlinks including inverses.



*LREC-COLING 2024*, pages 12053–12073
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

| Annotation scheme | TLink types | Events IAA | TLinks IAA | TLink type IAA | IAA Metric | Corpus genre | Num. events | Num. TLinks |
|---|---|---|---|---|---|---|---|---|
| TimeML (Pustejovsky et al., 2005, 2010a) | 13 | 0.78 | n/a | 0.55 | AvgPnR | news | 7,935 | 3,481 |
| TempEval-1 (Verhagen et al., 2007, 2009) | 6 | n/a | n/a | 0.47 | Cohen $\kappa$ | news | 7,935 | 2,002 |
| TempEval-3 (UzZaman et al., 2012) | 13 | 0.87 | n/a | n/a | F1 | web | 11,145 | 11,098 |
| THYME-TimeML (Styler et al., 2014) | 5 | 0.79 | 0.50 | 0.50 | Krippend-orff $\alpha$ | clinical | 15,769 | 7,935 |
| Temporal Dependency Structure (Kolomiyets et al., 2012; Bethard et al., 2012) | 6 | 0.86 | 0.82 | 0.7 | | fables | 1,233 | 1,139 |
| MATRES (Ning et al., 2018) | 4 | 0.85 | n/a | 0.84[1] | Cohen $\kappa$ | news | 6,099 | 13,577[2] |
| RED (O'Gorman et al., 2016; Ikuta et al., 2014) | 4 | 0.86 | 0.73 | 0.18-0.54 | F1 | news | 8,731 | 4,969 |
| TimeBank-Dense (Cassidy et al., 2014) | 6 | n/a | n/a | 0.56-0.64 | Cohen $\kappa$ | news | 1,729 | 12,715 |
| NewsReader (Minard et al., 2016; van Erp et al., 2015) | 13 | 0.68 | n/a | n/a | Dice's coef. | news | 2,096 | 1,789 |
| Araki et al. (Araki et al., 2018) | 2 | 0.80 (F1) | n/a | 0.11-0.14 | Fleiss $\kappa$ | simple wiki | 5,397 | 2,833 |
| CaTeRS (Mostafazadeh et al., 2016) | 4 | 0.91 | n/a | 0.51 | Fleiss $\kappa$ | stories | 2,708 | 2,715 |
| UDS-T (Vashishtha et al., 2019) | 2 | n/a | 0.67 | n/a | Spearman | web | 32,302 | 70,368 |
| TDG (Yao et al., 2020) | 4 | 0.79 | 0.52-0.85 | 0.85-0.91 | F1 | wiki | 14,974 | 28,350 |
| MAVEN-ERE (Wang et al., 2022) | 6 | n/a | 0.678 | n/a | Cohen $\kappa$ | wiki | 103,193 | 1,216,217 |

[1] Both coefficients of agreement are reported for two expert annotators who annotated a small portion of data (about 100 events and 400 relations).
[2] Since the initial release MATRES was extended to include the entire TempEval3 dataset (only verbal events). We cite the numbers for the newer, extended version available at https://github.com/qiangning/MATRES.

Table 1: Statistics reported in the current temporal annotation projects for English.

annotation schemes. The results of annotation in either framework can be represented in ISO-TimeML *format* (Pustejovsky et al., 2010a) encoded as as a collection of TLINKS between event pairs.

## 2. Related work

To the best of our knowledge, all current proposals for temporal annotation are based on the event-pair framework. Within that framework, there are different annotation schemes that have been applied to different text corpora. A summary of major available resources is presented in Table 1, which shows that the task of annotating event order is not characterized by high agreement, and there is no real consensus even on what agreement metric to use. The reported IAA for identifying events tends to be considerably higher than IAA for either establishing TLINKS, or for their type.

A fundamental problem for temporal annotation is that a complete set of temporal relations in a text would be quadratic on the number of events in that text, and estblishing them all would be prohibitively labor-intensive. Therefore most of existing work limit the scope of the task: only annotating TLINKS in the same or adjacent sentences (Verhagen et al., 2007, 2010; UzZaman et al., 2012; Minard et al., 2016), limiting the scope to a specific construction (Bethard et al., 2007). Another line of work focuses on trying to infer the missing TLINKS via transitive closure (Setzer and Gaizauskas, 2001; Verhagen, 2005; Mani et al., 2006). However, this process is not conflict-free (Verhagen, 2005), and the current methods to produce full temporal graphs from sparse annotations are not very successful (Ocal et al., 2022a). A key problem is that the existing annotations often suffice only to construct local event chains, but there is not enough information to connect them (Chambers and Jurafsky, 2008).

In addition to laboriousness, establishing the set of all possible TLINKS is difficult because human readers do not even infer all of these relations for every text they read. Much of this information is underspecified, and if the annotators are forced to infer it, their agreement would not be high. The chief solution for underspecification has been to either allow sparse annotation, to introduce additional restrictions to avoid annotating non-actual events (Bethard et al., 2012) or, more recently, place them on separate axes (Ning et al., 2018).

We contribute a new annotation framework, which replaces individual event pairs with a holistic view of the narrative represented as a timeline. This solves the density problem: as shown in Figure 1, a timeline contains all the information needed for ordering *all* event pairs. It also enables a novel solution to the underspecification problem: we incorporate vagueness in the event type definitions that have different timeline visualisations (see §3.1). Finally, it is more aligned with the natural human reading process (see Appendix A.)

Since we do not directly annotate TLINKS, but a structure from which they can be unambiguously inferred, our approach resembles the annotation of temporal dependency graphs and trees (Kolomiyets et al., 2012; Zhang and Xue, 2018, 2019; Yao et al., 2020), where the annotators establish temporal relations as child-parent relation-

ship in a dependency tree. However, that approach has to assume a single parent-child relation, and the annotation process still requires considering individual pairs of events or events with temporal expression, while we allow for event clusters (§3.4). The dependency structure is also less amenable to express vagueness and underspecification than our timeline-based proposal. Furthermore, temporal dependency trees may be more temporally indeterminate than the TimeML annotations (Ocal and Finlayson, 2020).

A number of previous projects used timeline-like representations (Verhagen et al., 2006; Kolomiyets et al., 2012; Do et al., 2012; Caselli and Vossen, 2016, 2017), but only as a representation of the final result: the annotation itself was still based on event pairs. Vashishtha et al. (2019) proposed a framework where the annotators work with only two adjacent sentences to create a mini-timeline of the events in those two sentences. This enables crowdsourcing, but necessarily limits the annotation to adjacent sentences (and only a subset of those, in practice). Most recently, Wang et al. (2022) stated that they developed and used a timeline-based annotation scheme to improve annotation density, but provided no further details, tools or the guidelines with which this was achieved.

## 3. NarrativeTime framework

Temporal annotation is usually performed in two stages: (1) identification of events, and (2) their temporal ordering. NarrativeTime focuses on (2): as shown in Table 1, detection of events is an easier task with a relatively high IAA. We do not introduce anything new here, and consider events as "anything that happens or occurs" (Pustejovsky et al., 2003a), expressed as verbs, nominals, adjectives/-participles, or phrases. States also count as events. Since we re-annotate TimeBank data, we use the original event annotations.

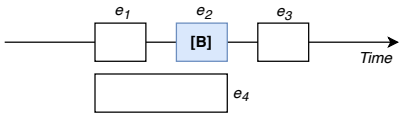
### 3.1. Event types

Most current annotation schemes adopt a model of temporal relations based on interval algebra (Allen, 1984), where the start and endpoints of 2 events form 13 possible relations: Before/After, Immediately Before/After, Overlap/Is overlapped, ends/is ended on, starts/is started on, during, and identity. But full tracking of all the event start/endpoints is psychologically unrealistic.

We propose integrating some temporal order information in event definitions rather than leaving it all to tlinks. The annotators need to be able to focus on the start, end, or the ongoing phase of an event, or any combination thereof that is salient in the context, and leave out the underspecified parts. This idea owes a lot to the huge body of linguistic work on verb aspect and event structure (Dowty, 1986; Pustejovsky, 1991; Moens and Steedman, 1988; Smith, 1997), verb classes (Vendler, 1957; Levin, 1993; Chipman et al., 2017), and particularly the geometric event phase representations by Croft (2012). To the best of our knowledge, this is the first attempt to merge aspectual and event order[3] information in a single annotation unit (in TimeML they are separate).

To achieve this, NarrativeTime distinguishes between bounded, unbounded and partially bounded[4] events, defined as follows.

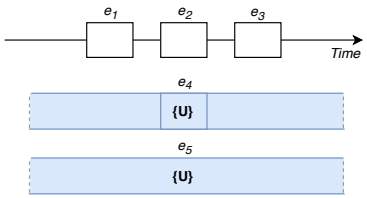
**Bounded events [B]** are events (of any nature and duration) that are known to start roughly after the end of the nearest other event on the timeline, and they end before the next event starts (with or without a temporal gap). In the example in Figure 2, the event of Mary packing ($e_2$) is "bounded" by the events of her coming ($e_1$) and leaving ($e_3$). John working is also a bounded event, the duration of which spans $e_1$:$e_2$. The start of $e_1$ and the end of $e_3$ are "bounded" by the start/end of the story.




Example: *John started working$_4$ when Mary came in$_1$, and stopped when she packed$_2$ and left$_3$ for New York.*

Figure 2: Bounded events

**Unbounded events {U}** are events (of any nature and duration), of which the exact start and end points are not known, but they are known to overlap with some other event on the timeline, and possibly (in an underspecified way) with its neighbors.




Example: *Mary went$_1$ to the coffee shop and found$_2$ John there. He was working$_4$ on his lifelong project$_5$. She left$_3$.*

Figure 3: Unbounded events

[3] Reimers et al. (2016) proposed distinguishing between "single-day" and "multi-day" events, but this was to enable anchoring to temporal expressions rather than to annotate event order.

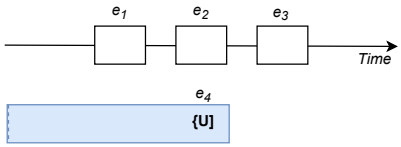
[4] We hope that the linguist reader will excuse our redefining "boundedness", an established term in Aktionsart literature.

In the example in Figure 3, the event of John working ($e_4$) started at an underspecified point, possibly before Mary started walking to the coffee shop ($e_1$). We also don't know when he stops working; maybe immediately after Mary's leaving ($e_3$), and maybe hours later. The only thing we know for sure is that he was working when Mary saw him ($e_2$), and this is what {U} events encode in NarrativeTime. The temporal location of [B] event $e_2$ is used as the temporal "center" of the {U} event $e_4$.

A big advantage of this definition of unbounded events is that it singles out the cases where the exact temporal order is underspecified, but some inference about relations of events surrounding the anchor [B] event and the {U} event may be possible based on the world knowledge.[5]

We also define a special case of "permanent" unbounded events, represented in this example by event $e_5$ (John's lifelong project). This is an event that occurs throughout the narrative, and likely also beyond it. Such events are also of {U} type, but they are not "centered" on any particular slot on the timeline. We use this mechanism to account for relatively permanent characteristics of characters and entities, which are unlikely to change in the course of the narrative (e.g. "John is dark-haired"), and generic events (e.g. "people like coffee").

**Partially bounded events [U}, {U]** are a combination of the two above types, used when one endpoint of an event is known, and the other endpoint is underspecified. Figure 4 illustrates an event bounded on its right endpoint, and unbounded on the left. The event of Mary calling John ($e_2$) is "anchoring" the {U] type event of John's working $e_4$, which lasts during[6] her calling him and for some underspecified time prior to that. He was probably working while she was walking, but that is in the sphere of inference based on world knowledge.




Example: *Mary walked$_1$ across the garden. She called$_2$ for John. He stopped working$_4$, and they left$_3$ together.*

Figure 4: Partially bounded events

## 3.2. Factuality

Another source of uncertainty in the temporal annotation is events for which that is not clear, such as future events, negated events, conditionals, modals, comparisons, and figures of speech. Ning et al. (2018) address that problem by placing events with different realis status on different timelines, so as to not annotate underdefined relations.

Our solution is based on the possible-worlds approach: all such events are treated as real events on the timeline for the purposes of establishing temporal order. For example, if a text mentions that John didn't send a birthday present to his mother, this non-event is in fact an event with a certain timeline location. To account for the realis status, we introduce a simplified version of FactBank (Saurí and Pustejovsky, 2009) factuality markup, which combines the axes of negation (happened/didn't happen) and certainty (did happen/maybe happened).

This gives us four possible values for factuality. Since most events in narrative texts are of the "happened" type, in NarrativeTime they are left unmarked for factuality. The other types can be manually specified in the "factuality" column in the annotation interface (Figure 6) with the following simple text markers: **"-"** for "didn't/won't happen", **"m"** for "maybe happened/will happen", and **"m-"** for "maybe didn't/won't happen".

## 3.3. Timeline branches

The relations between all events on a coherent timeline can be expressed with the bounded/unbounded event mechanism (§3.1). But often there is not enough information for such a timeline. In the example in Figure 5, we know that John read the book before watching the movie, but it is not clear if he read it before or after coming to Boston.

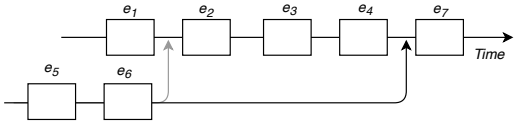
NarrativeTime handles such cases by creating a branch on the main timeline. A branch is defined as a mini-timeline, linked with a before/after relation to some point on the main timeline. In the example in Figure 5, one such candidate attachment point is the movie visit. The events on the branch happen in parallel to the events in the corresponding section of the main timeline, and are in a Vague relation

[5] In this example, our intuition is that it didn't take Mary long to get to the coffee shop, so John was probably working while she was getting there. Such guesses are not in the scope of event order annotation, but there are relevant efforts to collect data about possible event durations (Vashishtha et al., 2019) and commonsense reasoning (Qin et al., 2021; Zhou et al., 2019). Given that we have some extra mechanism for reasoning about likely event durations, NarrativeTime annotation could tell where such reasoning would be warranted. Leeuwenberg and Moens (2020) take the opposite approach and directly elicit annotations of the upper/lower bounds of events.

[6] NarrativeTime annotators are free to choose the level of granularity of event order. For example, we might interpret John stopping to work as something that happens *after* Mary calling him: e.g. if we know that John is not someone to spring up instantly, or if it is a crime story where the exact order matters. But the interval is so small that in most cases these events could be considered roughly simultaneous. NarrativeTime can accommodate either interpretation, depending on annotator instructions or the saliency of the event order.

| | Cluster type | Description | Example |
|---|---|---|---|
| [B] | Clusters of roughly-simultaneous bounded events. | A [B] event can denote a single bounded event or a cluster, where the events are either roughly-simultaneous, or their order does not matter for the current narrative. | *John called, texted and left voicemails for Mary incessantly* |
| [C] | Clusters of consecutive events. | Narratives often contain mini-scripts, or combinations of cause/effect, enabling/enabled events that could only happen in that order. | *John brushed his teeth and got dressed*<br>*John woke up and thought of Mary.* |
| {U} | Clusters of unbounded events. | Narratives often contain descriptive sequences, where the temporal information for all named features is the same. Hence they can all be annotated as a single {U} event. | *John was a short, fat man with a red face and a bald patch* |

Table 2: Event cluster types in NarrativeTime




Example: *John came* $_{e_1}$ *back to Boston. (...) He bought* $_{e_2}$ *a ticket, had* $_{e_3}$ *a coffee and headed* $_{e_4}$ *to the cinema. He had already read* $_{e_5}$ *the book and he liked* $_{e_6}$ *it. The movie started* $_{e_7}$*.*

Figure 5: Branching timelines in NarrativeTime

to them. Since it takes longer to read a book than to get to a movie theater, we could infer[7] that the book was read before the movie-related sequence.

There are two types of branches: for event(s) happening at some time before a given point (marked **<]** ), or after a given point (marked [**>** ).

### 3.4. Event clusters

Psychologists established that texts that are pre-chunked in semantically coherent segments are easier to process (Frase and Schwartz, 1979; O'Shea and Sindelar, 1983; Rajendran et al., 2013). For dynamic situations in the narratives, we hypothesize that "semantic coherence" is best explained in terms of scripts/frames. For example, the sentence "John woke up, brushed his teeth, got dressed, went to the office, and proposed to Mary", is likely to be remembered as 2 events rather than 5: the morning-routine event and the proposal event.

NarrativeTime leverages this feature of human reading comprehension by encouraging the annotators to think in terms of event clusters and not single events. In particular, we define the types of event clusters that are presented in Table 2. Annotators with different cognitive styles could choose to process a particular sequence as a cluster or individual events – but they would still produce annotations that are equivalent in terms of event order sequence on the timeline.

### 3.5. Anchoring of temporal expressions

NarrativeTime follows Pustejovsky et al. (2005) in defining temporal expressions (timex). We make no contribution in this area, and use the pre-existing timex annotations of TimeBank in our case study. What NarrativeTime does improve is their linking with events: annotators only need to include any temporal expressions in the event spans which they anchor, so the spans function as temporal containers (Pustejovsky and Stubbs, 2011). No further action is needed for event-timex links.

For example, if [*John met Mary on* ***Monday***] is chosen as the event span, then the meeting event would be anchored to Monday. If a cluster of simultaneous events is in the same span as a timex, then all of them are anchored to that timex.This approach echoes treating temporal expressions as event arguments, which reportedly reduces the annotation effort by 85% as compared to TimeBank-Dense (Reimers et al., 2016). If a timex applies to several consecutive events (e.g. from timeline position 2 to 5), it is possible to create a separate timex span and specify its duration as an interval (e.g. 2:5). If for some reason an event and its timex cannot be in the same span, the same position on the timeline can be assigned for them individually.

### 3.6. Annotation workflow

NarrativeTime comes with a new open-source web-based annotation tool. The interface for annotating event order[8] is shown in Figure 6.

An annotation is created by choosing the event type ([B] by default), highlighting some span in the text, and either accepting the auto-populated values of time, branch, and factuality, or manually editing them in the annotation table. By default,

[7] Whether to perform this extra reasoning step turned out to be a big source of disagreement. We experimented with forcing the annotators to attach branches simply where they were mentioned, but this extra reasoning is a part of natural reading process, hard to suppress consistently. We believe this is one of the reasons why temporal annotation generally suffers from low IAA.

[8] In this study, we used pre-annotated events and event coreference information from the original TimeBank, but our annotation tool also has a basic interface for annotating events and their coreference.

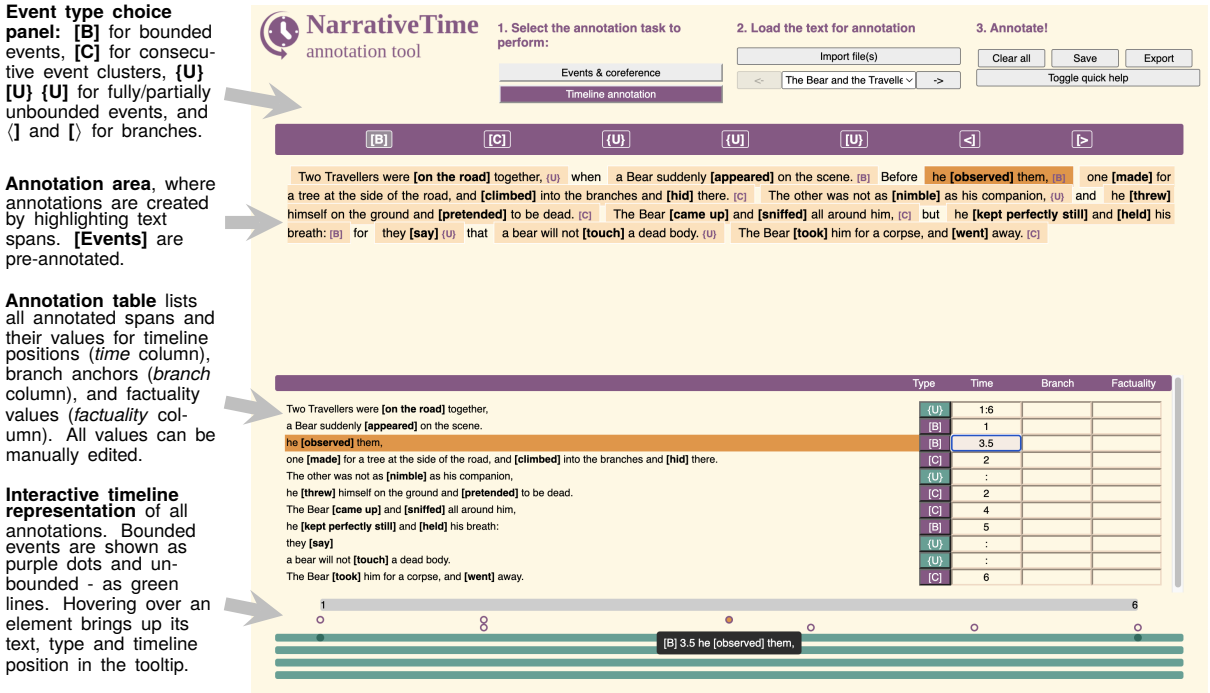



Figure 6: NarrativeTime annotation interface

new annotations are bounded, actual events on the main timeline, at the position after the previous highest one (e.g. if the timeline ends at position 2, then a new [B] event will be placed at 3).

This workflow minimizes the number of clicks: the best case scenario is that the annotators only need to read the text, highlighting events in chronological order. That will auto-populate the timeline integers serving as timeline position indicators. To "move" an event to another timeline position only its *time* value needs to be edited. This way it is easy to insert new events without changing existing annotations: e.g. if there are events at positions 1 and 2, a new event can be placed between them by setting its *time* value to 1.5. The type of an existing annotation can also be changed (by clicking on the type button in the annotation table).

It is possible to annotate the order of individual events by highlighting them individually, but, as shown in Figure 6, the tool also allows annotating multi-event spans, interpreted as clusters of bounded, unbounded, or bounded consecutive events (§3.4). This both saves annotation effort, and allows to leverage the natural chunking-during-reading strategies of the annotators.

A limitation of the current annotation tool is that each event span is associated with only one point on the timeline. However, in practice we have not yet encountered cases in which the same event should map to non-adjacent points.

The NarrativeTime tool uses its own format for representing timelines, and is accompanied by a script for conversion to the ISO-standard TimeML format (Pustejovsky et al., 2010b) (with the addition of the factuality annotations in the format similar to FactBank (Saurí and Pustejovsky, 2009)). Examples of both formats and more details are available in the repository. We use 5 classic TimeML relations (BEFORE/AFTER, INCLUDES/IS_INCLUDED, SIMULTANEOUS), as well as VAGUE (Verhagen et al., 2007) and OVERLAP (Verhagen et al., 2007).

## 4. Evaluation of annotation

**TimeBankNT corpus.** In scope of this work, we re-annotate 36 documents of the TimeBank corpus which were also used in TimeBank-Dense (Cassidy et al., 2014), MATRES (Ning et al., 2018) and TD-Discourse (Naik et al., 2019). This enables direct comparison between the different methodologies.

Two first authors of this paper were both the annotators and the main developers of the guidelines, which underwent many rounds of revision (based on annotating news and fiction texts and discussing cases of disagreement). After that, we created two full annotations for each of 36 TimeBank-Dense documents. The final corpus contains 1,715 original event and 289 timex annotations, to which we added 2 independent NarrativeTime annotation sets. Each set contains 1,715 factuality annotations, 79,001 event-event TLINKS, 23,979 event-timex TLINKS, and 1,770 timex-timex TLINKS. Statistical information as a table is available in the Appendix D (Table 8). See Figure 9 for the distribution of TLINKS labels.

| | Event Type | Event Order | Factuality | Branching |
|---|---|---|---|---|
| Agreement Rate | 0.88 | 0.75 | 0.93 | 0.92 |
| Cohen's $\kappa$ | 0.62 | 0.68 | 0.84 | 0.68 |
| Krippendorff's $\alpha$[10] | 0.62 | 0.68 | 0.83 | 0.68[11] |

Table 3: NarrativeTimeinter-annotator agreement.

**Inter-annotator agreement.** We compute four types of IAA: event type, factuality, branching and event order. For event types, we compare if both annotators chose the same type (e.g., [U}) for the given event. For event order, we convert[9] NarrativeTime annotation to TimeML format, using the approach described in appendix B, and compare all event-event and event-timex TLINKs for all 7 relation types in our conversion scheme. This tests both timeline and event type annotation, as event relations depend on both. For factuality, we compare whether a given event has the same factuality annotation (incl. the default empty value, which corresponds to non-negated actual events). For branching, we check if both annotators placed the event to a branch instead of the main timeline. The results are shown in Table 3.

Our results for event type, event order and branching could be described as "substantial agreement", and for factuality – as "perfect agreement" (Landis and Koch, 1977; Artstein and Poesio, 2008). The prior results for temporal order annotation (with IAA estimated as Cohen $\kappa$ or Krippendorff $\alpha$) are in the range of 0.47-0.84 (see Table 1). However, the direct comparison with annotation of event pairs is not fair to NarrativeTime, because we are solving a more difficult task: NarrativeTime annotators have to guarantee that a given annotation is consistent with all other existing annotations, which is not the case in pairwise approach.

Figure 7 shows that by far the most frequent event type was bounded events [B] (1446 spans where both annotators selected this type), followed by {U} (116) and [U} (88). The most confusion between event types was between [B] and {U} (59), and [B] and [U} (87). For the temporal relations, Figure 8 shows that a big contributor to confusion is the Vague relation, as well as simultaneous vs includes/is_included and simultaneous vs before/after.

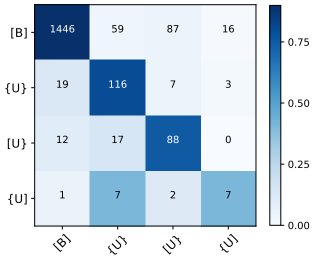



Figure 7: Event type confusion matrix

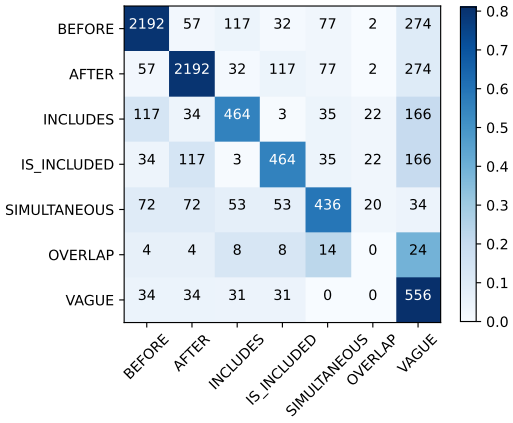



Figure 8: TLINK relation type confusion matrix

To explore the causes of disagreement, we performed a full qualitative evaluation of 6 documents with varying IAA values. We found that only 8% of disagreements are due to mistakes, and the majority would be more appropriately described as "human label variation" (Plank, 2022; Uma et al., 2021). The common causes include differences in the granularity of interpretation (8%), in the perception of the event endpoints (12%), interpreting events as states vs actions (20%), attribution of events to a timeline position (22%), and interpreting event clusters as consecutive vs roughly-simultaneous (30%). Our results suggest that higher IAA may not be achievable in full temporal annotation of realistic newswire texts. See Appendix C for more details.

**Annotation density.** Table 4 shows the base statistics and TLINK-to-event ratio for the densest, to our knowledge, currently available English resources with temporal annotation. Among them, the densest expert-annotated resources are TimeBank-Dense (Cassidy et al., 2014) and the recent MAVEN-ERE (Wang et al., 2022). Our solution

[9]Since the clustering mechanism of NarrativeTime allows for different span annotations with equivalent timelines (§3.4), computing agreement directly on span annotations would both the temporal order and the individual differences in chunking strategies.

[10]Both agreement coefficients reported here, $\kappa$ and $\alpha$, are values computed on the entire dataset, not averages of values for each document.

[11]The binary nature of branching (the decision whether or not to place an event on a branch) makes the data distribution naturally skewed as the majority of events are on the main timeline. Computing agreement coefficient such as $\alpha$ or $\kappa$ on skewed distribution results here in lower agreement as represented by these coefficients, which ultimately creates the relatively big gap between the agreement rate (0.98) and agreement coefficients (0.68) (Di Eugenio and Glass, 2004; Paun et al., 2022).

| Project | Events | Timexes | tlinks | Ratio |
|---|---|---|---|---|
| TempEval-3 UDS-T | 32,302 | – | 70,368 | 2.20 |
| TimeBank-Dense | 1,729 | 289 | 12,715 | 7.40 |
| TDDiscourse | 1,729 | 289 | 6,150 | 3.05 |
| MATRES | 6,099 | 1,955 | 13,577 | 1.69 |
| TDT-Crd | 2,691 | 1,414 | 4,105 | 1.0 |
| TDG | 14,974 | 2,485 | 28,350 | 1.62 |
| Event Storyline | 7,275 | 1,297 | 4,017 | 0.47 |
| MAVEN-ERE | 103,193 | 25,843 | 1,216,217 | 9.43 |
| NarrativeTime | 1,715 | 289 | 102,313 | **51.05** |

Table 4: Density of tlinks backed by manual annotation in the densest temporal annotation resources currently available for English. The density is computed as total number of tlinks (without inverses), divided by (number of events + number of timexes). See Table 9 for comparison with more resources.

is 5 times denser than than the previous densest solution, MAVEN-ERE. Table 4 reports only the number of event-event tlinks without inverse relations; the total number of tlinks in TimeBankNT is 207,496 (for each annotator).

As discussed in §2, the sparsity problem with annotation based on event pairs is usually addressed by trying to infer the missing relations by transitive closure. With such inferred relations, the above-cited resources could be represented as much larger in terms of tlinks, but it would not be a fair comparison: our framework guarantees that the entire timeline is considered by the annotator, and hence all tlinks are backed by manual annotation. In temporal closure, they are only backed by the closure rules, and because of incomplete, conflicting, or missing annotations, the full temporal graph often cannot be constructed (Ocal et al., 2022a).

**Annotation speed.** This depends on the length of the text, and the complexity of temporal relations in it. A long stretch of text describing events that happen sequentially or roughly-simultaneously could be annotated with a single click. The speed also improves with annotator experience. At the end of the project, we could fully annotate an average TimeBank text in about 20-30 minutes.

## 5. Baseline results

**Methodology.** As a baseline model to estimate the difficulty of temporal relation classification based on NarrativeTime data, we develop a simple Transformer-based model. It consists of a LongT5[12] (Guo et al., 2021) encoder and a relation classification head. Our choice of LongT5 is motivated by its support for long documents (some of the annotated documents are as long as 2000 tokens), and its availability in different sizes (to investigate the effect of encoder size on performance).

We split the TimeBankNT corpus into the training set (30 documents) and test set (6 documents), and fine-tune[13] our system on the former. During training, we feed a whole TimeBank document into the encoder and then extract contextualized representations of each event and timex into a tensor $H \in \mathbb{R}^{[e \times h]}$. Then, we add a trainable bilinear form to predict relations between every pair of events as $H \cdot W \cdot H^T$, where $W \in \mathbb{R}^{[h \times r \times h]}$, $r$ is the number of relation types and $h$ is the hidden size of LongT5. We performed manual hyperparameter tuning of learning rate and weight decay for each encoder. After initial tuning, the variation of accuracy (within a single model) was at most 0.03. Final hyperparameters were: batch size `32`, learning rate `1e-4`, weight decay `0`, dropout `0.1`.

**Results.** The results are presented in Table 5. As basic baselines, we used both the most frequent class and a simple rule that assigns events as After if they occur later in the text in relation to other events and Before otherwise. Human results are for one annotator vs. the other.

Even the best model only reaches F1 of 0.31, which shows that the task is challenging – but there is a large gap with the human performance, despite the human label variation. One challenge is that the temporal relations between nearby events (within 10 tokens) are the hardest to predict ($F_1$ of 0.19 vs 0.31 for all events). Another issue is imbalance in the distribution of relation data.[14] At the same time, our simple "later is after" heuristic baseline achieves only 30% accuracy, which shows that the temporal structure of these texts is indeed complex. We also find that the model does not rely excessively on either annotator (see Appendix E.)

**Suggestions for future work.** Since our texts are relatively long (up to 4K tokens), one direction for

[12]We do not present in-context learning with large language models (LLMs) as a baseline, since the high density of tlinks means that a single generation cannot produce all the thousands of relations that are typically present in a NarrativeTime document, and generating relations one by one is prohibitively expensive, especially via a paid API. One more issue is due to our reuse of TimeBank data: it is a very popular dataset present in many GitHub repositories, which makes it highly likely that popular LLMs had observed this data coupled with prior temporal annotations in pre-training. According to C4 search tool (`https://c4-search.apps.allenai.org/`), LongT5 was exposed to TimeBank texts, but we did not find TimeML annotations.

[13]We used a single A100 40Gb GPU, bf16 precision. The longest training run took about one hour.

[14]The Before/After relation covers ≈ 30% tlinks, Vague – ≈15%, Includes/Is_included and Simultaneous – ≈ 8%, and Overlap – only 0.2%

| | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Most frequent class | 0.30 | 0.04 | 0.15 | 0.07 |
| "Later is after" heuristic | 0.30 | 0.09 | 0.14 | 0.11 |
| LongT5 Base (114M) | 0.44 | 0.32 | 0.29 | 0.29 |
| LongT5 Large (349M) | 0.45 | 0.35 | 0.28 | 0.29 |
| LongT5 XL (1253M) | 0.47 | 0.34 | 0.31 | 0.31 |
| Human performance | 0.73 | 0.58 | 0.59 | 0.57 |

Table 5: Modeling results. Precision, recall, and F1 are macro-averaged over relation types. 'Human performance' refers to one annotator vs another.

follow-up work is long context models like LLaMA2 (Touvron et al., 2023) or Mixtral 8x7B (Jiang et al., 2024), and methods that significantly reduce memory requirements for large texts, such as multi-query attention (Shazeer, 2019). Our dataset provides a testbed for evaluating such models on long-distance relations, with the caveat of likely training data contamination by earlier TimeBank versions.

Our baseline uses a standard approach to relation prediction through a learnable bidirectional form $W_{out}$. Similar to BERT-like approaches, we replace the language modeling head with a new matrix of learnable parameters $W_{out}$. But this significantly differs from text-to-text approach prevalent in modern NLP, and the naïve approach of predicting all pairs of relations in text-to-text fashion (e.g., event2 is after event1) does not scale to the number of relations in our dataset. Developing alternative approaches and possibly modeling NarrativeTime annotation explicitly could improve the results, if we find better ways to communicate ordinal relations between annotations (event1 timestamp = 3, event2 timestamp=4) to the model.

## 6. Future work

The NarrativeTime improvements in temporal annotation density and handling of underspecification open up several exciting prospects for future work.

**More data with dense temporal annotation.** By enabling dense temporal annotation at a fraction of the cost of full manual annotation with traditional event pairs, NarrativeTime provides a means to create new resources for training ML models and more challenging benchmarks, in particular for long-distance temporal relations (Naik et al., 2019).

**Fine-grained vagueness.** One big problem with prior sparse approaches is being able to tell *why* no temporal relation is assigned between a given pair of events: did the annotator just not consider it, or considered it and decided that no relation exists, or that multiple relations are possible (Chambers et al., 2014)? NarrativeTime solves this problem by (a) ensuring that annotator does explicitly consider every possible relation by putting everything on a timeline, (b) providing three mechanisms for handling different cases of underspecification: timeline branches, unbounded events, factuality values.

Since NarrativeTime explicitly distinguishes between temporal order underspecification due to unbounded events, different timeline branches, or factuality, these cases can now be targeted for additional commonsense reasoning annotation and inference (Zhou et al., 2019). For example, in a sentence *John woke up, went to work, got off the bus, came to the office, stopped his podcast.* we don't know exactly when he started listening to the podcast, but we know it probably did include the bus time because people often listen to podcasts when they commute. Given NarrativeTime annotation, we would be able to tell when the model should try to reason about likely event duration.

**The death of the "gold standard"?** This work showed a significant amount of genuine variation in temporal annotation Appendix C, which reinforces the need to move away from the traditional "gold standard" approach to temporal annotation (Plank, 2022). Rather than trying to adjudicate such cases, we need to start modeling the possible interpretations by different people. We release TimeBankNT version of TimeBank-Dense corpus, fully double-annotated, and we hope that NarrativeTime framework would enable more such resources.

**Generalization to other domains and languages.** While this study focuses on news, we also experimented with fiction, encyclopedia, and fables. More systematic work is needed, but we were able to annotate all the phenomena we encountered in these domains with the proposed framework. We have not tested the annotation tool with other languages, but it depends on white space tokenization, which can be added in pre-processing even for languages like Japanese. The auto-numbering of bounded spans on the timeline works in the order in which they are selected by the annotator, so it should work even if the annotator reads the text right-to-left.

## 7. Conclusion

We present NarrativeTime, a new framework for temporal annotation that is based on a timeline representation of the whole text, rather than the order of individual event pairs. NarrativeTime achieves IAA comparable or superior to the prior art on news texts, but it offers the densest possible annotation, three mechanisms for handling underspecification, and support for a more natural reading process. We contribute NarrativeTime guidelines, open source tools for annotation and conversion to the standard TimeML format, as well as TimeBankNT corpus: the densest TimeBank, with 36 texts each annotated by two expert annotators.

## Ethics statement

All annotation work on TimeBankNT was performed by the authors of the submission. The news articles for annotation come from the original TimeBank corpus (Pustejovsky et al., 2005, 2010a), and also used in TimeBank-Dense (Cassidy et al., 2014), MATRES (Ning et al., 2018) and TDDiscourse (Naik et al., 2019). We do not foresee any additional risks created by this project.

While this submission focuses on validating the proposed NarrativeTime framework by reannotating a well-studied English resource, its broader impacts could include faster and easier creation of resources with dense temporal annotation for other domains and languages.

## Data and code availability

The code for the annotation tool, conversion to TimeML format, annotation guidelines, and all annotated data (in both NarrativeTime and TimeML formats) are available in the project repository[15] under MIT license.

## Acknowledgements

We would like to show our appreciation to the anonymous reviewers for their valuable feedback which allowed us to improve this work. This work was funded in part by an NSF CAREER award (IIS-1652742) to Anna Rumshisky.

## Bibliographical References

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

Jun Araki, Lamana Mulaffer, Arun Pandian, Yukari Yamakawa, Kemal Oflazer, and Teruko Mitamura. 2018. Interoperable Annotation of Events and Event Relations across Domains. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 10–20. ACL.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

S. Bethard, J. H. Martin, and S. Klingenstein. 2007. Timelines from Text: Identification of Syntactic Temporal Relations. In *International Conference on Semantic Computing (ICSC 2007)*, pages 11–18.

Steven Bethard, Oleksandr Kolomiyets, and Marie-Francine Moens. 2012. Annotating Story Timelines as Temporal Dependency Structures. In *Language Resources and Evaluation Conference*, pages 2721–2726.

Antonija Blaži Ostojić. 2023. Reading comprehension processes: A review based on theoretical models and research methodology. *Hrvatska revija za rehabilitacijska istraživanja*, 59(1):122–143.

Tommaso Caselli and Piek Vossen. 2016. The Storyline Annotation and Representation Scheme (StaR): A Proposal. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 67–72. ACL.

Tommaso Caselli and Piek Vossen. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86. ACL.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An Annotation Framework for Dense Event Ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506. ACL.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Nathanael Chambers and Daniel Jurafsky. 2008. Jointly Combining Implicit Constraints Improves Temporal Ordering. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 698–706, Honolulu, Hawaii. Association for Computational Linguistics.

Susan E. F. Chipman, Martha Palmer, Claire Bonial, and Jena Hwang. 2017. VerbNet: Capturing English verb behavior, meaning, and usage. In Susan E. F. Chipman, editor, *The Oxford Handbook of Cognitive Science*. Oxford University Press.

Kiel Christianson. 2016. When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *The Quarterly Journal of Experimental Psychology*, 69(5):817–828.

Berry Claus. 2012. Processing Narrative Texts: Melting Frozen Time? *Constraints in Discourse 3: Representing and Inferring Discourse Structure*, 223:17.

[15] https://github.com/text-machine-lab/nt

Marta Coll-Florit and Silvia P. Gennari. 2011. Time in language: Event duration in language comprehension. *Cognitive Psychology*, 62(1):41–79.

William Croft. 2012. *Verbs: Aspect and Causal Structure*. Oxford Linguistics. Oxford University Press, Oxford [England]; New York.

Barbara Di Eugenio and Michael Glass. 2004. Squibs and discussions: The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Quang Xuan Do, Wei Lu, and Dan Roth. 2012. Joint Inference for Event Timeline Construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 677–687, Stroudsburg, PA, USA. ACL.

David R. Dowty. 1986. The effects of aspectual class on the temporal structure of discourse: Semantics or pragmatics? *Linguistics and philosophy*, 9(1):37–61.

Christine Farag, Vanessa Troiani, Michael Bonner, Chivon Powers, Brian Avants, James Gee, and Murray Grossman. 2010. Hierarchical Organization of Scripts: Converging Evidence from fMRI and Frontotemporal Degeneration. *Cerebral Cortex*, 20(10):2453–2463.

Fernanda Ferreira, Paul E Engelhardt, Paul Engelhardt, and Manon W Jones. 2009. Good Enough Language Processing: A Satisficing Approach. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 31:413–418.

Lawrence T. Frase and Barry J. Schwartz. 1979. Typographical cues that facilitate comprehension. *Journal of Educational Psychology*, 71(2):197–206.

Mandy Guo, Joshua Ainslie, David C. Uthus, Santiago Ontañón, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. In *NAACL-HLT*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting Narrative Timelines as Temporal Dependency Structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97. ACL.

K. Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*. Content Analysis: An Introduction to Its Methodology. Sage.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Artuur Leeuwenberg and Marie-Francine Moens. 2020. Towards Extracting Absolute Event Timelines From English Clinical Reports. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2710–2719.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

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

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. In *Language Resources and Evaluation Conference*, pages 4417–4422.

Jennifer B. Misyak, Morten H. Christiansen, and J. Bruce Tomblin. 2010. On-Line Individual Differences in Statistical Learning Predict Language Processing. *Frontiers in Psychology*, 1.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational linguistics*, 14(2):15–28.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and Temporal Relation

Scheme for Semantic Annotation of Event Structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61. ACL.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A Dataset for Discourse-Level Temporal Ordering of Events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A Multi-Axis Annotation Scheme for Event Temporal Relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. ACL.

Mustafa Ocal and Mark Finlayson. 2020. Evaluating Information Loss in Temporal Dependency Trees. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2148–2156, Marseille, France. European Language Resources Association.

Mustafa Ocal, Adrian Perez, Antonela Radas, and Mark Finlayson. 2022a. Holistic Evaluation of Automatic TimeML Annotators. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1444–1453, Marseille, France. European Language Resources Association.

Mustafa Ocal, Antonela Radas, Jared Hummer, Karine Megerdoomian, and Mark Finlayson. 2022b. A comprehensive evaluation and correction of the TimeBank corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2919–2927, Marseille, France. European Language Resources Association.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Lawrence J. O'Shea and Paul T. Sindelar. 1983. The Effects of Segmenting Written Discourse on the Reading Comprehension of Low- and High-Performance Readers. *Reading Research Quarterly*, 18(4):458–465.

Silviu Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical methods for annotation analysis*. Springer Nature, Cham, Switzerland.

Barbara Plank. 2022. The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *EMNLP*. arXiv.

James Pustejovsky. 1991. The syntax of event structure. *Cognition*, 41(1):47–81.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003a. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The TIMEBANK Corpus. In *Proceedings of Corpus Linguistics*, pages 647–656.

James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. *The language of time: A reader*, pages 545–557.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010a. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA).

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010b. Iso-timeml: An international standard for semantic annotation. In *LREC*.

James Pustejovsky and Amber Stubbs. 2011. Increasing Informativeness in Temporal Annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. ACL.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal Commonsense Reasoning in Dialog. *arXiv:2106.04571 [cs]*.

Jamie M. Quinn, Richard K. Wagner, Yaacov Petscher, and Danielle Lopez. 2015. Developmental Relations Between Vocabulary Knowledge and Reading Comprehension: A Latent Change Score Modeling Study. *Child Development*, 86(1):159–175.

Dhevi J. Rajendran, Andrew T. Duchowski, Pilar Orero, Juan Martínez, and Pablo Romero-Fresco. 2013. Effects of text chunking on subtitling: A quantitative and qualitative examination. *Perspectives*, 21(1):5–21.

Keith Rayner and Erik D. Reichle. 2010. Models of the Reading Process. *Wiley interdisciplinary reviews. Cognitive science*, 1(6):787–799.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal Anchoring of Events for the TimeBank Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2195–2204. ACL.

Roser Saurí and James Pustejovsky. 2009. FactBank: A Corpus Annotated with Event Factuality. *Language Resources and Evaluation*, 43(3):227–268.

Elizabeth R. Schotter, Randy Tran, and Keith Rayner. 2014. Don't Believe What You Read (Only Once): Comprehension Is Supported by Regressions During Reading. *Psychological Science*, 25(6):1218–1226.

Alix Seigneuric, Marie-France Ehrlich, Jane V. Oakhill, and Nicola M. Yuill. 2000. Working memory resources and children's reading comprehension. *Reading and Writing*, 13(1):81–103.

Andrea Setzer and Robert Gaizauskas. 2001. A Pilot Study On Annotating Temporal Relations In Text. In *Proceedings of the ACL 2001 Workshop on Temporal and Spatial Information Processing*.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need.

Wayne L. Shebilske and L. Starling Reid. 1979. Reading Eye Movements, Macro-structure and Comprehension Processes. In Paul A. Kolers, Merald E. Wrolstad, and Herman Bouma, editors, *Processing of Visible Language*, Nato Conference Series, pages 97–110. Springer US, Boston, MA.

Carlota S. Smith. 1997. *The Parameter of Aspect*, 2. ed edition. Number 43 in Studies in Linguistics and Philosophy. Kluwer Academic Publ, Dordrecht.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2012. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*.

Elke van der Meer, Reinhard Beyer, Bertram Heinze, and Isolde Badel. 2002. Temporal order relations in language comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 28(4):770–779.

Marieke van Erp, Piek Vossen, Rodrigo Agerri, Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2015. Annotated Data, version 2. Technical Report D3-3-2, VU Amsterdam.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-Grained Temporal Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. ACL.

Zeno Vendler. 1957. Verbs and Times. *The Philosophical Review*, 66(2):143.

M. Verhagen, R. Knippen, I. Mani, and J. Pustejovsky. 2006. Annotation of Temporal Relations with Tango. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA).

Marc Verhagen. 2005. Temporal closure in an annotation environment. *Language Resources and Evaluation*, 39(2/3):211–241.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 Task 15: TempEval Temporal Relation Identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80. ACL.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval challenge: Identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden, 15-16 July 2010. ACL.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction.

Jiarui Yao, Haoling Qiu, Bonan Min, and Nianwen Xue. 2020. Annotating Temporal Dependency Graphs via Crowdsourcing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5368–5380, Online. Association for Computational Linguistics.

Yuchen Zhang and Nianwen Xue. 2018. Structured Interpretation of Temporal Relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuchen Zhang and Nianwen Xue. 2019. Acquiring Structured Temporal Representation via Crowdsourcing: A Feasibility Study. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 178–185, Minneapolis, Minnesota. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "Going on a vacation" takes longer than "Going for a walk": A Study of Temporal Commonsense Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3361–3367, Hong Kong, China. Association for Computational Linguistics.

Rolf A. Zwaan. 2016. Situation models, mental simulations, and abstract concepts in discourse comprehension. *Psychonomic Bulletin & Review*, 23(4):1028–1034.

## A. Why event pairs are problematic: motivation in psychology

The exact mechanisms of reading comprehension are still debated (Rayner and Reichle, 2010; Blaži Ostojić, 2023), but there are good reasons to believe that we gradually build a mental model of the whole narrative (van der Meer et al., 2002; Zwaan, 2016). This model has a directional representation of time and temporal distance between events, and is built correctly even if the text is not organized chronologically, e.g. if there are flashbacks (Claus, 2012).

We also know that texts pre-chunked in semantically coherent segments are easier to process (Frase and Schwartz, 1979; O'Shea and Sindelar, 1983; Rajendran et al., 2013). For dynamic situations, "semantic coherence" is best explained in terms of scripts/frames, mental representations of stereotypical complex activities. They have internal organization, with possibly complex sub-elements that can be managed without losing track of the overall goal of the script (Farag et al., 2010).

The process of constructing a mental model of a narrative is likely to be subject to the same on-line constraints[16] as the rest of language processing. This brings into play the "good-enough processing" (Christianson, 2016; Ferreira et al., 2009). Not all temporal relations *can* be inferred, since the writers focus on advancing their story in an engaging way rather than spelling out all the details. The readers also have limited time and attention, and focus on salient developments with the characters, often ignoring the details. This is the fundamental reason for the underspecification problem in temporal annotation.

Counter-intuitively, readers do *not* save effort by looking at each segment only once: we regress as needed (Schotter et al., 2014), even across sentence boundaries (Shebilske and Reid, 1979). This suggests that during reading a good-enough representation of the narrative is constructed, with the readers anticipating the developments (Coll-Florit and Gennari, 2011) and filling the most glaring gaps with their world knowledge. The variation is particularly notable with regards to the length of durative events (Coll-Florit and Gennari, 2011). This would explain the relatively low inter-annotator agreement observed in previous temporal annotation projects.

If the above view of reading comprehension is correct, it is the opposite of the process required from annotators in a schema based on event pairs. The annotators are explicitly asked about the temporal order of two events, which may or may not be in the category of events that were salient enough in the discourse to be easily order-able. Furthermore, there is no allowance for the fact that underspecified relations are not just "vague": if they are salient enough, their order *will* be inferred, but that interpretation may well be different for different annotators, since they draw on their own world knowledge (see appendix C for examples of such cases).

[16] Reading comprehension in particular is influenced by the working memory capacity (Seigneuric et al., 2000), vocabulary proficiency (Quinn et al., 2015), and even individual differences in statistical learning (Misyak et al., 2010).

## B. Post-processing

Given our new definitions of event types, we developed a new representation for NarrativeTime annotation that is used internally in the annotation tool. This is a simple json-based format containing the indices of pre-annotated timexes, events, and their coreference chains, as well as the indices and timeline positions, types, actuality, and branch annotations for the timeline annotations. A small example of this format is shown in Listing 1; see the project repository for more details.

The internal format allows for underspecification in temporal relations through the NarrativeTime mechanisms (branches, factuality, and unbounded events). However, the current standard for representing temporal information is based on event-event or event-time pairs, specifically, TimeML-ISO (Pustejovsky et al., 2010a), and this is what most existing applications expect. Hence we also provide a tool for converting the NarrativeTime annotation to the more familiar TimeML TLINKS (see the project repository for details). We opted to use 5 classic TimeML relations (BEFORE/AFTER, INCLUDES/IS_INCLUDED, SIMULTANEOUS), as well as VAGUE (Verhagen et al., 2007) and OVERLAP (Verhagen et al., 2007) Without the inverse relations (BEFORE/AFTER, INCLUDES/IS_INCLUDED), the set could be reduced to 5. This mapping is external and auxiliary to NarrativeTime, and other mappings could also be developed.

Listing 2 shows the data from Listing 1 represented in with TimeML (for text and TLink tags) and FactBank (for FACT_VALUE tags) style. This is a small example with only 4 events and 1 timex, and we do not show the possible inverse relations (which would double the overall amount of TLINKS), but the explicit enumeration of all possible TLINKS still looks more verbose, and harder to fix errors in.

The format conversion also involves significant conceptual trade-offs, since it requires a mapping between NarrativeTime format, which represents the vague relations with the combination of unbounded events and branching mechanism, and the classical TimeML relations. Our choices are shown in Table 6, with examples of overlapping and non-overlapping temporal intervals indicating the timeline positions for different combinations of

event types.

The first column (the case of two bounded events **[][]**) is simple and corresponds to the classical TimeML relations, but the cases involving unbounded events (**[}, {]** and **{}**) are more difficult. We opted to map to VAGUE (empty cell in the table) all cases where more than one relation could theoretically be possible: for example, an unbounded event at position {3} necessarily INCLUDES a bounded event at position [3], but its position with respect to another unbounded event at position {3} could be either SIMULTANEOUS or OVERLAP, depending on the exact edges of the two events (underspecified by definition, could only be resolved with case-by-case commonsense reasoning or by providing more contextual information).

As evident from Table 6, this means losing information, since NARRATIVETIME format can express the difference between the vagueness on both or one end[17] of an unbounded event. It also does not allow for differentiation between vagueness due to unboundedness and branching. Future work could explore learning/predicting temporal information directly from NARRATIVETIME representation, or developing more fine-grained types of VAGUE for the classical TimeML representation.

For the events in the branches, their relations with events/timexes on the main timeline is determined by their anchor position and their direction. For example, if a branch is anchored at position 3 and goes into the future, its events are AFTER any main timeline events prior to 3, and VAGUE with the events after position 3 (since they exist in a parallel world, so to speak).

## C. Qualitative Analysis

We manually analyzed 6 documents (4,336 TLINKS)[18] to identify the cases where annotators' interpretations differ, resulting in label variation.[19] The analysis was performed on the original timeline-based annotations, rather than on the TimeML conversion. This allowed us to compare the annotation without losing any information due to conversion. We identified 5 main types of variation between the annotators, listed in Table 7.

We observe that the biggest single source of label variation stems from the decision to cluster several events together as roughly simultaneous, or explicitly mark their order (see CATEGORY 1 in Table 7). This is not actually disagreement, but expected variation in chunking strategies between annotators, which can still produce temporal annotations equivalent in terms of TLINKS

We further notice that for some events, there may be more than one plausible temporal interpretation: this source of variation corresponds to CATEGORY 2 in Table 7. Consider the "issues" in the example (a). Since they concern a crime, one interpretation is that the issues existed since the crime was committed. Another interpretation is that the issues concern the court case. Since that set is not exactly the same as all issues concerning the crime, in that case, they only exist since the court case.

Note that this kind of difference in temporal perception may also result in varying, yet equally acceptable, annotations of the event factuality. For instance, "find" in example (b) can be interpreted as negated event in the past (i.e., "didn't happen"), or a potential event in the future (i.e., "maybe will happen"). All examples in this category rely heavily on the annotator interpretation, which can differ due to individual differences, cultural background, etc.

An almost equally common reason for label variation is "state vs action" (CATEGORY 3 in the table): one annotator puts more focus on the underlying action, while the other focuses on the resulting state. This results in seeing the same event as either a bounded event positioned in the past or a partially unbounded event (state) continuing into the future. For instance, "decapitated" ($e_2$) from the example in Table 7 can be interpreted as a bounded event [B] in the past when the action of decapitation took place, or as the state [U} resulting from that action, which started at the same moment as the action, but then continued indefinitely into the future.

The differences in the perceived scope of the event (CATEGORY 4) are usually related to attitude verbs, such as "think" or "believe", which in news texts usually come in official statements. One possible interpretation is that the attitude is held at the moment of speech, in which case they would be annotated as bounded events ([B]). But it is also plausible that the attitude is held for some time before/after expressing that attitude; in that case they would be annotated as unbounded {U} events "centered" at the moment of speech.

Finally, we observe some differences due to

[17] In the pairwise approach, the partial unboundedness could be partially implemented by introducing additional START_ON and END_BY relations, but this would require an additional TLINK to specify the VAGUE relation at the other end of the interval. If such an event is "centered" on several other events rather than one, even more annotation would be needed.

[18] The agreement on TLINKS for the documents sampled for the qualitative analysis ranges from Krippendorff's $\alpha$=0.47 (one of the lowest) to $\alpha$=0.85 (one of the highest). Choosing documents with varying agreement allows us to analyze both cases where the annotators tend to interpret the timeline uniformly, and cases where their interpretations are more likely to differ.

[19] Here we use the term "variation" rather than "disagreement" following a recent proposal in Plank (2022), since disagreement implies that both interpretations cannot hold. Cases where none or only one interpretation is plausible were classified as mistakes.

```
{
    # text id
    "id":"sample",

    # space-tokenized text
    "text":"John ordered a new bike for his summer trip , but his order got lost .",

    # "spring" timex annotation: [start token, end token]
    "timex": {"0": [7, 7]},

    # similarly structured event annotations ("ordered", "used", "order", "lost")
    "events":{"0":[1,1],"1":[8,8],"2":[12,12],"3":[14,14]},

    # coreference chain between "ordered" (token 1) and "order" (token 12)
    "event_coreference":{"1":[12]},

    # events in coreference chains are unmarked for annotation, except for the first mention
    "invisible_events": [12],

    # timeline annotation
    "event_order":{
        "0":{"span":[0,4],"type":0,"time":"1","factuality":"","branch":""},
        "1":{"span":[6,8],"type":0,"time":"3","factuality":"m","branch":""},
        "2":{"span":[11,14],"type":0,"time":"2","factuality":"","branch":""}}
    # "span": [start token, end token] for the annotated span
    # "type": the span types (0=[B], 1=[C], 3={U}, 4=[U}, 5={U])
    # "time": the timeline position of the annotated span
    # "factuality": factuality annotation
    # "branch": the timeline attachment point of a branch + its type
}
```

Listing 1: NarrativeTime native format example

```
<?xml version="1.0" encoding="utf-8"?>

<TimeML>

John <EVENT eid="0">ordered</EVENT>a new bike for his <TIMEX3 tid="t0">summer</TIMEX3><EVENT eid="1">trip</
     EVENT>, but his <EVENT eid="2">order</EVENT>got <EVENT eid="3">lost</EVENT> .

<MAKEINSTANCE eiid="ei0" eventID="0"/>
<MAKEINSTANCE eiid="ei1" eventID="1"/>
<MAKEINSTANCE eiid="ei2" eventID="2"/>
<MAKEINSTANCE eiid="ei3" eventID="3"/>

<FACT_VALUE eiid="ei0" fvid="1" value="CT+"/>
<FACT_VALUE eiid="ei1" fvid="2" value="PS+"/>
<FACT_VALUE eiid="ei3" fvid="3" value="CT+"/>
<FACT_VALUE eiid="ei2" fvid="4" value="CT+"/>

<TLINK lid="1" eventInstanceID="ei0" relType="BEFORE" relatedToEventInstance="ei1"/>
<TLINK lid="2" eventInstanceID="ei0" relType="BEFORE" relatedToTime="t0"/>
<TLINK lid="3" eventInstanceID="ei0" relType="BEFORE" relatedToEventInstance="ei3"/>
<TLINK lid="4" eventInstanceID="ei0" relType="SIMULTANEOUS" relatedToEventInstance="ei2"/>
<TLINK lid="5" eventInstanceID="ei1" relType="AFTER" relatedToEventInstance="ei0"/>
<TLINK lid="6" eventInstanceID="ei1" relType="SIMULTANEOUS" relatedToTime="t0"/>
<TLINK lid="7" eventInstanceID="ei1" relType="AFTER" relatedToEventInstance="ei3"/>
<TLINK lid="8" eventInstanceID="ei1" relType="AFTER" relatedToEventInstance="ei2"/>
<TLINK lid="9" timeID="t0" relType="AFTER" relatedToEventInstance="ei0"/>
<TLINK lid="10" timeID="t0" relType="SIMULTANEOUS" relatedToEventInstance="ei1"/>
<TLINK lid="11" timeID="t0" relType="AFTER" relatedToEventInstance="ei3"/>
<TLINK lid="12" timeID="t0" relType="AFTER" relatedToEventInstance="ei2"/>
<TLINK lid="13" eventInstanceID="ei3" relType="AFTER" relatedToEventInstance="ei0"/>
<TLINK lid="14" eventInstanceID="ei3" relType="BEFORE" relatedToEventInstance="ei1"/>
<TLINK lid="15" eventInstanceID="ei3" relType="BEFORE" relatedToTime="t0"/>
<TLINK lid="16" eventInstanceID="ei3" relType="AFTER" relatedToEventInstance="ei2"/>
<TLINK lid="17" eventInstanceID="ei2" relType="SIMULTANEOUS" relatedToEventInstance="ei0"/>
<TLINK lid="18" eventInstanceID="ei2" relType="BEFORE" relatedToEventInstance="ei1"/>
<TLINK lid="19" eventInstanceID="ei2" relType="BEFORE" relatedToTime="t0"/>
<TLINK lid="20" eventInstanceID="ei2" relType="BEFORE" relatedToEventInstance="ei3"/>

</TimeML>
```

Listing 2: Listing 1 data represented in TimeML and FactBank style

different granularity of annotation of unbounded events (CATEGORY 5). One annotator could interpret an event as a generic/permanent state (unbounded event without a temporal position, encoded as {:}), while another could attribute it to a specific period in time + underspecified periods before/after (encoded as {x} or {x:y}).

Unavoidably, we also find some mistakes, mostly (but not only) due to annotating an event and a timex under the same span. While this annotation

| $e_1$ TIME | $e_2$ TIME | $[e_1]\,[e_2]$ | $\{e_1\}\,\{e_2\}$ | $[e_1\}\,[e_2\}$ | $[e_1]\,\{e_2]$ | $[e_1]\,\{e_2\}$ | $\{e_1\}\,[e_2]$ |
|---|---|---|---|---|---|---|---|
| 1:3 | 4:6 | BEFORE | | | | | |
| 4:6 | 1:3 | AFTER | | | | | |
| 1:6 | 3:4 | INCLUDES | | | | | INCLUDES |
| 3:4 | 1:6 | IS_INCLUDED | | | | IS_INCLUDED | |
| 1:4 | 3:6 | OVERLAP | | | | | |
| 3:6 | 1:4 | OVERLAP | | | | | |
| 1:3 | 1:3 | SIMULTANEOUS | | | | IS_INCLUDED | INCLUDES |

| $e_1$ TIME | $e_2$ TIME | $[e_1]\,[e_2]$ | $[e_1]\,\{e_2\}$ | $[e_1\}\,[e_2\}$ | $[e_1]\,\{e_2]$ | $\{e_1]\,[e_2]$ | $\{e_1\}\,[e_2\}$ |
|---|---|---|---|---|---|---|---|
| 1:3 | 4:6 | BEFORE | BEFORE | | | BEFORE | |
| 4:6 | 1:3 | AFTER | | AFTER | AFTER | | |
| 1:6 | 3:4 | INCLUDES | | INCLUDES | | INCLUDES | |
| 3:4 | 1:6 | IS_INCLUDED | IS_INCLUDED | | IS_INCLUDED | | |
| 1:4 | 3:6 | OVERLAP | OVERLAP | | | OVERLAP | |
| 3:6 | 1:4 | OVERLAP | | OVERLAP | OVERLAP | | |
| 1:3 | 1:3 | SIMULTANEOUS | IS_INCLUDED | INCLUDES | IS_INCLUDED | INCLUDES | INCLUDES |

| $e_1$ TIME | $e_2$ TIME | $[e_1]\,[e_2]$ | $[e_1\}\,\{e_2\}$ | $\{e_1\}\,\{e_2\}$ | $\{e_1]\,\{e_2\}$ | $\{e_1]\,[e_2]$ | $[e_1\}\,\{e_2\}$ |
|---|---|---|---|---|---|---|---|
| 1:3 | 4:6 | BEFORE | | | | BEFORE | |
| 4:6 | 1:3 | AFTER | | | | | AFTER |
| 1:6 | 3:4 | INCLUDES | | | | | |
| 3:4 | 1:6 | IS_INCLUDED | | | | | |
| 1:4 | 3:6 | OVERLAP | | | | OVERLAP | |
| 3:6 | 1:4 | OVERLAP | | | | | OVERLAP |
| 1:3 | 1:3 | SIMULTANEOUS | IS_INCLUDED | INCLUDES | IS_INCLUDED | OVERLAP | OVERLAP |

Table 6: Mapping of interval relations to TimeML relations. The first two columns show examples of overlapping and non-overlapping temporal intervals indicating timeline positions of events (a single-value position X is equivalent to X:X interval, e.g. 3:3.) The remaining columns show different combinations of event types with these intervals. Empty cells indicate the VAGUE relation.

is not necessarily problematic, it may lead to errors when there is another event placed before or after the given event that also shares the same time (e.g., both events happen on the same day of the week). Overall, we notice that about 8% of the difference in annotations can be attributed to mistakes of one or both annotators. This compares to 13% errors in *all* TLINKS in TimeBank 1.2 (an improved version of TimeBank 1.1), reported by Ocal et al. (2022b).[20]

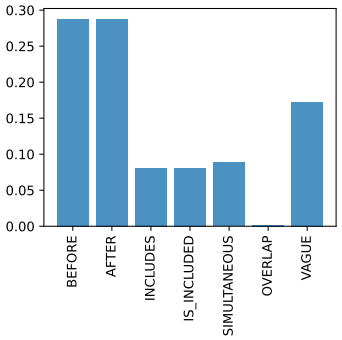



Figure 9: Relation type distribution

## D. Supplementary analysis

### D.1. Supplementary statistics

Table 8 shows the overall statistics for the annotated corpus in a table format, and Figure 9 presents the distribution of different types of TLINKS. Table 9 presents a comparison to a wider range of other resources in terms of density of TLINKS, complementary to the shorter Table 4. Figure 10 shows the confusion matrix for the event span types selected by the two annotators (complementing the confusion matrix for TLINKS in Figure 8).

### D.2. The use of NARRATIVETIME-specific annotation mechanisms

As described in §3, NARRATIVETIME proposes three mechanisms for handling underspecification: unbounded and partially bounded event type, branch-

[20] Note that these values (8% and 13%) are not directly comparable. In case of NARRATIVETIME, the 8% refers to the 8% of "disagreement" found in the 6 analyzed texts, while in the case of the TimeBank 1.2 the 13% refers to the 13% of *all* TLINKS (not only disagreement) in the texts analyzed in Ocal et al. (2022b).

| | Category | Description | Example | % |
|---|---|---|---|---|
| 1. | consecutive vs roughly-simultaneous | While one annotator groups the events together as roughly simultaneous, the other annotates the order explicitly. | No one was hurt, but firefighters [***ordered***]$e_1$ the [***evacuation***]$e_2$ of nearby homes and said they'll monitor the shifting ground. | 30% |
| | | Timeline: **[B] vs [B][B]** on the same temporal position | [***ordered***]$e_1$ and [***evacuation***]$e_2$ – consecutive *or* roughly simultaneous | |
| 2. | different positions on the timeline | The annotators differ in the way they interpret the event and its temporal position, but both interpretations are plausible. Note that this may also lead to different, yet equally acceptable, annotations of factuality (e.g., interpreting an event as one that did not happen in the past or as one that may happen in the future). | (a) Now the ninth US circuit court of appeals has ruled that the original appeal was flawed since it brought up [***issues***]$e_1$ that had not been raised before. (b) The police and prosecutors said they had identified different suspects in six of the cases and had yet to [***find***]$e_1$ any pattern linking the killings or the victims, several of whom were believed to be prostitutes. | 22% |
| | | Timeline: **[B] vs [B]** on different temporal positions | [***issues***]$e_1$ – when crime was committed *or* when they were brought up<br>[***find***]$e_1$ – past negated event or future possible event | |
| 3. | state vs action | While one annotator interprets the event as a state that begins at a certain point and lasts through a portion of the story (partially bounded), the other annotates it as a bounded event with the focus being on the action rather than the resulting state. | Kidnappers kept their promise to kill a store owner they took hostage and police found the man's [***dismembered***]$e_1$ and [***decapitated***]$e_2$ body Friday [***wrapped***]$e_3$ in plastic garbage bags. | 20% |
| | | Timeline: **[B] vs [U]/{U]** anchored on the same temporal position | [***dismembered***]$e_1$, [***decapitated***]$e_2$, [***wrapped***]$e_3$ – from certain point in the past (state) *or* at certain point in the past (action) | |
| 4. | bounded vs centered unbounded | While one annotator marks an event as bounded, the other treats it as an unbounded event "centered" at the same point as the bounded event in the other annotation.This difference in interpretation is common for attitude verbs such as "think," "hope." "believe." | And I [***hope***]$e_1$ that, whatever happens today, that our relationships with Russia will continue to be productive and constructive and strong, because that's very important to the future of our peoples. | 12% |
| | | Timeline: **[B] vs {U}** on the same temporal position | [***hope***]$e_1$ – at the given moment (bounded) *or* overlapping with neighboring events (centered unbounded) | |
| 5. | granularity | Annotators' interpretations differ in their level of granularity. These are usually cases where one annotator annotates an unbounded event as permanent/generic, and another annotator adds a "center" to that event. | There have been no [***arrests***]$e_1$ in any of the slayings. | 8% |
| | | Timeline: **{:} vs {U}/{U1:U2}** or {U1:U2} vs {U1:U2} with a wider interval | [***arrests***]$e_1$ – generally, in the whole story (unbounded) *or* up to the moment of the utterance (centered unbounded) | |
| 6. | mistakes | Any mistake due to honest lapses of judgment. Most mistakes can be attributed to accidentally marking two events or an event and a timex under the same span when that interpretation is impossible or results in other inconsistencies (e.g., marking another event that also relates to the same timex as before or after the event which is already annotated as simultaneous to the times). | "I haven't seen a pattern yet," [***said***]$e_1$ Patricia Hurt, the Essex County prosecutor, who [***created***]$e_2$ the task force on Tuesday.<br><br>One annotator accidentally groups [***said***]$e_1$ and [***created***]$e_2$ under one span. | 8% |

Table 7: Reasons for label variation between the annotators.

| | |
|---|---|
| Texts | 36 |
| Events | 1,715 |
| Timexes | 289 |
| Event-event TLINKs | 79,001 |
| Event-timex TLINKs | 23,979 |
| Timex-timex TLINKs | 1,770 |
| Factuality annotations | 1,715 |

Table 8: TimeBankNT corpus statistics

ing, and factuality. The substantial agreement on event types and branching, and perfect agreement on factuality (Table 3) provides evidence that the guidelines were sufficiently clear, and the annotators made use of these mechanisms in similar ways.

A key innovation in NarrativeTime framework is that it enables the annotation of event clusters (§3.4), rather than just individual events, which makes it possible to annotate multiple temporal relations at once. At the same time, whether to use this mechanism is up to the annotator, and it is certainly possible to produce equivalent timelines with different chunking strategies.

Figure 11 shows the overall distribution of events in the spans highlighted by both annotators: while the majority of annotations contain only one event,

| Project | Events | Timexes | tlinks | Ratio |
|---|---|---|---|---|
| TempEval-3 (UzZaman et al., 2012) | 11,145 | 2,078 | 11,096 | 0.84 |
| UDS-T (Vashishtha et al., 2019) | 32,302 | – | 70,368 | 2.20 |
| TimeBank-Dense (Cassidy et al., 2014) | 1,729 | 289 | 12,715 | 7.40 |
| TDDiscourse (Naik et al., 2019) | 1,729[1] | 289 | 6,150 | 3.05 |
| MATRES (Ning et al., 2018) | 6,099 | 1,955 | 13,577 | 1.69 |
| TDT-Crd (Zhang and Xue, 2019) | 2,691 | 1,414 | 4,105 | 1.0 |
| TDG (Yao et al., 2020) | 14,974 | 2,485 | 28,350 | 1.62 |
| Event Storyline (Caselli and Vossen, 2017) | 7,275 | 1,297 | 4,017 | 0.47 |
| MAVEN-ERE (Wang et al., 2022) | 103,193 | 25,843 | 1,216,217 | 9.43 |
| NarrativeTime | 1,715[2] | 289 | 102,313 | **51.05** |

Table 9: Density of tlinks backed by manual annotation in the current English resources. The density is computed as total number of tlinks (without inverses), divided by (number of events + number of timexes).

[1] TDDiscourse paper does not state the number of events; it is probably slightly smaller than in TimeBank-Dense, since their released data reference event IDs rather than event instance IDs. Only event-event tlinks seem to be annotated.
[2] The small discrepancy in event number between TimeBankNT and TimeBankDense is due to the fact that NarrativeTime annotation relies on event tokens rather than event instance tags (although we use the original TimeBank event instance id numbers in conversion).

almost one third of annotations contain two or more events. The distribution is very similar for the two annotators. This suggests that the span-based annotation is helpful for capturing temporal relations in the news genre, and we hypothesize that it could be even more useful for other genres with more temporally coherent chunks of text, such as descriptive paragraphs in fiction or historical narratives in encyclopedias.

## E. Additional results for baseline experiments

**Qualitative analysis of test documents.** For the test set, we select the same six documents, for which we had established through qualitative analysis (see appendix C) that the majority of disagreement cases are in fact human label variation. These documents vary in length, the number of events, and IAA (from $\alpha$=0.47 to $\alpha$=0.85). See Appendix E for additional analysis per test document.

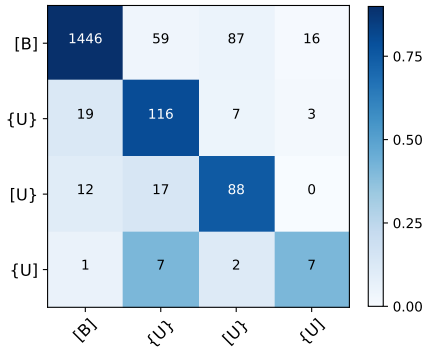



Figure 10: NarrativeTime event type confusion matrix for annotation

Looking at the per-document metrics (Figure 12) we observe that the system does not rely excessively on either of the annotators. In the case of `PRI19980115.2000.0186`, it could be related to the "consecutive vs roughly simultaneous" human label variation case (row 1 in Table 7). The `NYT19980402.0453` document is interesting because the IAA for it is low ($\alpha$=0.47), but the model's accuracy remains similar for both of the annotators.

The confusion matrix for our best model configuration (Figure 13) shows that the model overpredicts frequent Before and After relations (especially at the expense of Simultaneous), and almost never predicts the rare Overlap relations. Interestingly, the asymmetrical relations Before and After seem to be confused with another asymmetrical relations pair Includes and Is_included.

**Long-distance relations vs. short-distance relations** Since temporal relations between long-distance events are a distinctive feature of NarrativeTime, we perform an additional evaluation on events that are closer than ten words apart (i.e., roughly corresponding to adjacent sentences) vs. further than 100 words apart. Table 10 shows that both types of relations are hard to model, as the model makes more mistakes in these two classes than in general (Table 5). This suggests that medium-distance (10-100 words) relations are the simplest to predict. Low numbers on close-by event relations can be explained by the confusion between Simultaneous and Before/After (Figure 13), which in turn could be partly due not to errors, but to the label variation in consecutive vs roughly-simultaneous case (row 1 in Table 7).

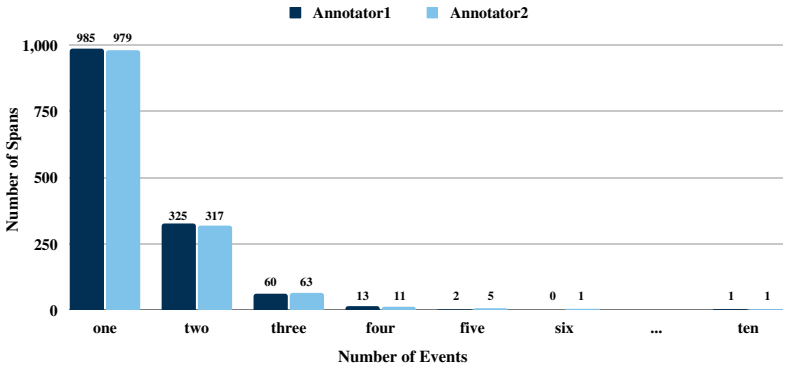



Figure 11: The number of span-based NARRATIVETIME timeline annotations with the number of events included in the spans. While the majority of spans contained only one event, almost one third encoded two or more events.

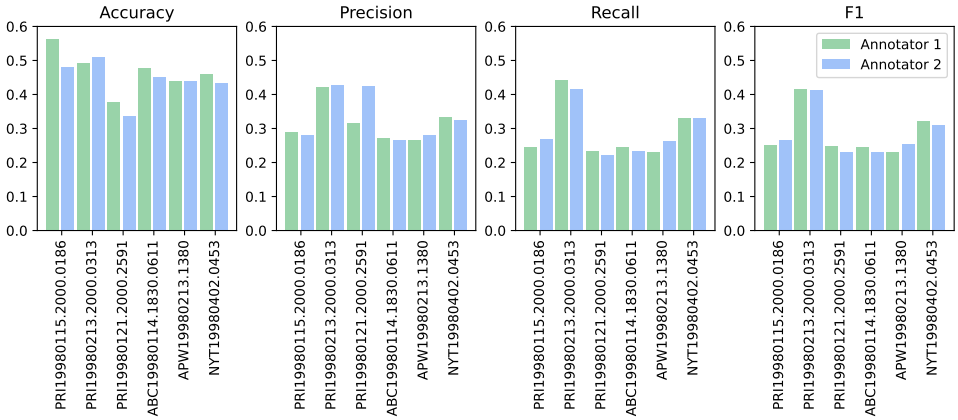



Figure 12: Per-document metrics

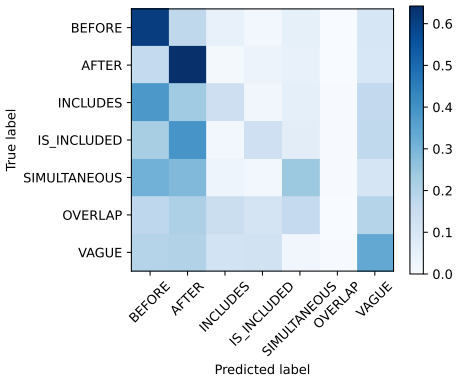



Figure 13: Relation prediction confusion matrix

| | ACCURACY | PRECISION | RECALL | F1 |
|---|---|---|---|---|
| All events | 0.47 | 0.34 | 0.31 | 0.31 |
| Nearby events | 0.27 | 0.17 | 0.23 | 0.19 |
| Far events | 0.41 | 0.32 | 0.28 | 0.29 |

Table 10: Short-distance and long-distance relations metrics.