7

The Richer Event Description Corpus for Event–Event Relations

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer

Abstract. A variety of approaches exist for annotating temporal and event information in text, but it has been difficult to compare and contrast these different corpora. The Richer Event Description (RED) corpus, as an ambitious annotation of temporal, causal, and coreference annotation, provides one point of comparison for discussing how different annotation decisions contribute to the timeline and causal chains which define a document. We present an overview of how different event corpora differ and present new methods for studying the impact of these temporal annotation decisions upon the resulting document timeline. This focuses on illuminating the contribution of three particular sources of information – event coreference, causal relations with temporal information, and long-distance temporal containment – to the actual timeline of a document. By studying the impact of specific annotation strategies and framing the RED annotation in the larger context of other event–event relation annotation corpora, we hope to provide a clearer sense of the current state of event annotation and of promising future directions for annotation.

7.1 Introduction

A crucial element of understanding any narrative or news report is an understanding of the timeline of events. This remains a difficult challenge for natural language processing (NLP) systems. However, although many different corpora have been build to handle temporal topics, there is still no consensus regarding the annotation of event-centric information such as temporal relations, causal chains, or event coreference. Though major advances have been made in formalizing markup languages for characterizing temporal relations (Pustejovsky, Castano et al., 2003; Pustejovsky, et al., 2010), there is less consensus on when to apply such labels, leading to large practical differences between corpora. There are also many temporally adjacent topics

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

that might provide temporal information when annotated, including event coreference (Bejan and Harabagiu, 2010; Lee et al., 2012), temporal relations (Pustejovsky, Hanks et al., 2003), modal or veridicality annotations (Saurí and Pustejovsky, 2009), annotations of the narrative structure or plot links of a documents (see Chapter 6), or annotations of causal relations and chains within a document.

As the number of possible event–event pairs explodes as the size of documents increases, it is generally considered impractical to annotate all possible relations; thus, event–event relations corpora are differentiated by the different strategies they use to define the subsets of those possible relations that should be considered during annotation. The Richer Event Description (RED; Ikuta et al., 2014; O'Gorman et al., 2016) corpus was created to be an ambitious annotation of a wide variety of temporal information. It exemplifies one kind of approach to that problem, in which annotation focuses on informative temporal containment relations, augmented by a range of local temporal order, causal relations, and event coreference annotations. This strategy used in RED differs from the recent trend toward dense locally windowed annotations of temporal order (Chambers et al., 2014; Ning et al., 2018; Vashishtha et al., 2019).

This work provides two primary contributions. It reframes the role of RED annotation in the larger context of different event–event relations corpora and presents a case study regarding how to make actual comparisons between such event–event relation corpora. It is hoped that this comparison will provide a launching point for the design of future approaches to temporal and event-related annotation corpora and better inform groups attempting to use this information. We suggest that evaluating corpora in terms of their actual contribution to an estimated timeline is an objective way of making these comparisons and may help to provide more empirical ways to evaluate event–event relations corpora.

7.2 A Comparison of Event Annotation Choices

Rather than discussing these characteristics in a vacuum, we compare the RED annotations to other annotations of event coreference and event–event relations, often having many layers of annotation applied to the same raw text. One of the earliest annotations of event information is that of TimeBank (Pustejovsky, Hanks et al., 2003), also augmented with fine-grained annotation of modality (Saurí and Pustejovsky, 2009), dense annotation of temporal relations (TB-Dense; Chambers et al., 2014), and causal relations (Mirza et al., 2014). The Universal Decompositional Semantics annotated temporal

relations (Vashishtha et al., 2019) over the English Web Treebank (EWT) corpus, which also had annotations of event factuality (Rudinger et al., 2018) and genericity (Govindarajan et al., 2019). A third major group of annotations is applied over the Event Coreference Bank (ECB+) corpus, containing both event coreference (Cybulska and Vossen, 2014) and the Event Storylines Corpus (Caselli and Vossen, 2017). We also compare to the CATERS (Mostafazadeh et al., 2016) and THYME (Styler et al., 2014) annotations, which apply temporal annotations over short texts and clinical notes, respectively. These corpora provide context to ground RED annotations within the larger space of possible annotation decisions.

7.2.1 Detection of Events and Their Features

Though there is wide disagreement regarding which event–event relations to annotate, there is much more agreement regarding what constitutes an event and the general types of features that those events might be annotated with.

Most event-based approaches handle all events and eventualities, as in the TimeML guidelines, which define events as "situations that happen or occur". This encompasses not only verbs but also timeline-relevant nouns, adjectives, and multiword expressions. Approaches vary in how much they annotate semantically light verbs such as auxiliaries and in how to handle agent-denoting nominalizations such as "prisoner" or "murderer".

A larger source of difference regarding what qualifies as an event comes from corpora that only annotate events when they fit into a particular ontology of interest. Such corpora (Song et al., 2015; Hong et al., 2016; Araki et al., 2018) thus would label a mention "sale" as an event only if the ontology cares about a category such as "transaction event". Wide-coverage event ontologies (such as Richer Event Ontology [REO]; see Chapter 2) may also capture all events in a document, and some approaches (Pustejovsky, Castano et al., 2003; Caselli and Vossen, 2017) annotate all events with event types by using coarse-grained event types, distinguishing broad categories such as mental state events.

Frameworks also differ in what features are annotated on those events (features sometimes applied during such an "event detection" phase), such as tense, modality, and genericity. Coarse-grained distinctions for these categories can often be applied using small sets of three to four labels. For example, the RED annotation provides coverage of the minimal set of distinctions for each such category – distinguishing modality/genericity with for features (ACTUAL, HYPOTHETICAL, GENERIC, and UNCERTAIN/HEDGED), along with positive or negative polarity. RED also marks the speech time relation – whether a

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

Table 7.1. *Different definitions events and tense and modality features*

Corpora	Event Definitions	Tense/ Event Time	Modality/ Polarity	Genericity	Event Types	Filter by Types
RED	all	✓	✓	binary		no
UDS-T	verbs+adj.	✓	✓	✓		no
TimeBank	all	✓	✓		some	no
THYME	all	✓	✓			no
ECB/ESC	all	✓	✓	✓	some	no
ERE/RPI	all		binary	binary	ACE	yes
CaTers	all				TRIPS	yes

RED = Richer Event Description; UDS-T = Universal Decompositional Semantics–Temporal; THYME = Temporal Histories of Your Medical Events; ESC = Event Storylines Corpus; ECB = Event Coreference Bank; ERE = Entity Relations and Events; RPI = Hong et al. (2016); CaTeRS = Causal and Temporal Relation Scheme.

given event is *before*, *after*, or *overlap*s with document time, including a special category of *before/overlap* for present prefect forms such as *"I have been writing for weeks"* – in a manner similar to other speech time annotations (Pustejovsky and Stubbs, 2011; Styler et al., 2014). Other annotation approaches use fine-grained or even continuous scales for aspect (Croft et al., 2016), event factuality (Saurí and Pustejovsky, 2009; Rudinger et al., 2018; Vigus et al., 2019), or genericity (Govindarajan et al., 2019).

RED annotates all eventualities, without any filtering based on ontology, so that all events in a document are annotated. This annotation therefore captures a superset of the events that one might need for a particular event–event relations task and with basic representations of most of the features one may want to find for those events.

7.2.2 When to Annotate Temporal Relations

The fundamental question to be resolved in temporal annotation is *when* to annotate the temporal relationship between two events. Because it is generally agreed that individually considering every possible event–event pair in a document is impractical, the fundamental question of temporal annotation has been how to define a set of constraints that can limit the number of relations considered to a tractable quantity.

One source of these constraints is to only annotate explicit temporal relations, signaled by temporal adpositions or adverbs (such as *before* or *since*) or other grammatical signals such as aspect marking. Such an approach is one

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

criteriona applied by RED and by other corpora such as TimeBank (Pustejovsky, Hanks et al., 2003) both label such relations and label the explicit signal itself. However, only marking such explicit relations would not be sufficient to capture the timeline of a document.

Beyond those clear-cut explicit relations, one set of strategies for annotating temporal relations is to have annotators view a full document and to select only a subset of salient temporal relations that the annotators decide to label. One approach to this is to emphasize only labeling informative temporal relations – those that would provide the most information about the underlying timeline (Pustejovsky and Stubbs, 2011) – by focusing upon hierarchical *containment* relations ("includes" in Allen interval relations), because if one can cluster events into larger macroevents, most event–event orderings can be inferred from simply knowing the order of those larger events. RED follows that narrative of containment-based annotations.

Such annotations have the advantage of capturing the most salient and useful labels but the disadvantage of providing a partial annotation of only these informative relations (Ning et al., 2019). Recently, many other annotations have defined more deterministic rules for which relations to consider – such as a window of adjacent sentences – and consider all event–event pairs that match such a deterministic heuristic (Chambers et al., 2014; Ning et al., 2018; Vashishtha et al., 2019). Such approaches have computational advantages, simplifying the hard task of detecting relationships into a simpler task of classifying them.

A third such strategy is to limit the space of possible relations using a temporal dependency approach (Kolomiyets et al., 2012; Zhang and Xue, 2018), where each event is only linked to a single other temporally related event. Zhang and Xue (2018) found the temporal link to another event that provides the most specific information about the current event.

An important factor in these different strategies is that they vary in how, if ever, they can capture long-distance temporal relationships. Explicit marking of temporal order is usually limited to adjacent sentences, and "dense" annotations of temporal order tend to have constrained windows of annotation.[1] However, the container-based approaches such as RED can mark relations over long distances, both between events and by marking containment between temporal expressions and contained events. RED also provides methods for marking long-distance temporal relations through an additional constraint discussed in the next section – adding temporal relations labels to causal

[1] One exception to this is Hong et al. (2016), which attempts a largely pairwise annotation of events between documents.

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

relationships, which are marked whenever an event clearly causes or enables another event. This provides a relatively limited set of contexts where a long-distance relation might be annotated, allowing annotations such as RED to tie together other unconnected sections of text.

7.2.3 Temporal Relations Covered

Despite the dramatic differences regarding when to annotate a temporal relation, there are not pursuantly dramatic differences regarding the actual inventories of temporal labels utilized by different annotation schemes. Though temporal inventories differ slightly from group to group, they naturally tend to be a subset of the Allen temporal interval relations (Allen, 1983). TimeML (Pustejovsky, Castano et al., 2003) established a set of approaches to annotate a subset of those intervals, and many later documents followed the same approach. Such annotations differ minimally in how they treat a small set of less frequent relations – whether to mark SIMULTANEOUS relations, to mark BEGINS-ON and ENDS-ON relations, and whether to encode aspectual – i.e., ALINK – relations. REO (see Chapter 2) has borrowed these basic temporal labels from RED in order to capture prototypical sequences of events for generic complex events.

Table 7.2 illustrates the temporal inventories for a number of these annotation frameworks. The most important deviation in this framework is not RED (which marks a conventional inventory of relations derived from TimeML) but rather Universal Decompositional Semantics–Temporal (UDS-T; Vashishtha et al., 2019), which relates events using a *continuous* representation of event

Table 7.2. *Temporal relation inventories, extending table from Mostafazadeh et al. (2016)*

Corpus	Before/ After	Meets	Overlap	Finishes/ End-on	Starts/ Begins-on	Contains/ Includes	Equals/ Identity
Allen	✓	✓	✓	✓	✓	✓	
RED	✓		✓	✓	✓	✓	✓
THYME	✓		✓	✓	✓	✓	✓
TimeML	✓	✓		✓	✓	✓	✓
CaTeRS	✓		✓			✓	✓
ESC	✓		✓			✓	

RED = Richer Event Description; THYME = Temporal History of Your Medical Events; CaTeRS = Causal and Temporal Relation Scheme; ESC = Event Storylines Corpus.

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

spans using two slider bars. Such an annotation can be reduced to the same information modeled in the other temporal relation inventories but adds useful continuous information about how the start and end points relate to each other, beyond what is captured by discrete temporal interval labels.

7.2.4 Event Coreference and Non-temporal Relations

RED augments temporal annotation with causal annotations (Ikuta et al., 2014), subtyping the temporal BEFORE and OVERLAP relations with casual counterparts into BEFORE/CAUSE, OVERLAP/CAUSE, BEFORE/PRECONDITION, and OVERLAP/PRECONDITION. Other annotations have also combined temporal and causal annotations (Mirza et al., 2014; Mostafazadeh et al., 2016).

Another class of nontemporal information is the subtyping of temporal CONTAINS relations into those that are purely temporal and those that express an event–subevent link – expressed in RED as CONTAINS-SUBEVENT. Being able to distinguish subevents can be important in the transition from pure temporal structure to an understanding of events and scripts. This is not unique to RED, because such links are annotated in Glavaš et al. (2014) and annotated within events of interest in Araki et al. (2014). However, the RED annotations are the largest such annotations of these subevent labels that we know of and may provide a useful starting point for script learning and characterization of larger events.

Finally, RED marks event coreference relations and partial coreference links (such as set/subset relations). Knowing whether two events are coreferent during a temporal information pass has useful implications for the larger representation of a document, because it allows annotators to avoid redundantly marking a given temporal relation multiple times in a document. RED also annotated set/member relations – such as between a class of events (as in a generalization) and individual instances. Hong et al. (2016) annotated not only event coreference and set/member but a long tail of other nontemporal event–event relations, such as subevent–subevent, member–member, and recurrence relations. The Event StoryLines Corpus (Caselli and Vossen, 2017) also provides explanatory PLOT LINK annotations, which similarly can be viewed as providing temporal information.

7.2.5 Actual Annotation Methods

There are two ongoing issues with event–event relation corpora that are fundamentally at odds: how to make annotations provide large amounts of scalable training data and how to provide extremely high-quality annotations.

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

The RED corpus provides one method of addressing the quality issue, by focusing on providing high-quality annotations through double annotation and adjudication of event detection and event feature annotations, only annotating temporal relations on top of the cleaned, adjudicated events. Such an approach minimizes the propagation of errors caused by low-level disagreements, such as the span of an event or whether to interpret a given event as generic or negated. By attempting to address such disagreements early in annotation, the RED annotations are as coherent as possible.

The RED approach differs from decompositional approaches, in which many different annotators provide annotations of event features or event–event relations with fine-grained or even continuous representations. Such approaches can be viewed as providing less coherent annotations (lacking such adjudication) but may offset that lack of coherence with richer information about the range of labels provided by different annotators.

7.2.6 Summarizing the Role of RED in Event Corpora

Tables 7.1, 7.2, and 7.3 show how different annotation approaches vary in their annotation decisions. One can see that more directly through example (1), which illustrates the events annotated in RED. One can see that all

Table 7.3. *Comparison of event–event relation annotations*

Corpora	Local Temp. Order	Long-distance Containment	Causal Relations	Plot Relations	Event Coreference
RED	salient	✓	✓		✓
TimeBank	salient	✓			
TB-Dense	all pairwise				
UDS-T	most pairwise				
Ning et al.	most pairwise				
ECB	plot-relevant		✓	✓	✓
CaTers	salient		✓		
THYME	salient	✓			

RED = Richer Event Description; TB-Dense = TimeBank Dense; UDS-T = Universal Decompositional Semantics–Temporal; ECB = Event Coreference Bank; THYME = Temporal Histories of Your Medical Events; CaTeRS = Causal and Temporal Relation Scheme.

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

eventualities – including pronouns such as *it* – are annotated as events. In RED, all such events also are annotated, in this case, as being BEFORE speech time and being of modality ACTUAL:

(1) Gadhafi's **[visit]** to Italy **[continued]** that process of **[emergence]** from international **[isolation]**. But **[it]** also drew **[protests]**, including at La Sapienza university, where Gadhafi was **[addressing]** a group of few hundred students.

We can see from such an example the kinds of relations that are captured as RED-style annotation through explicit and local temporal links such as aspectual marking and containment relations (CONTAINS-SUBEVENT):

	continued	CONTINUES	emergence
(2)	isolation	ENDS-ON	emergence
	visit	CONTAINS-SUBEVENT	addressing

The same sentences in RED would also receive event coreference and causal relations annotation for other events:

	it	OVERLAP/PRECONDITION	protests
(3)	visit	COREFERENCE	it
	visit	OVERLAP/CAUSES	continued

More dense annotations would get human judgments for other relations beyond these, including pairs such as (visit, emergence), (continued, isolation), or (isolation, protests).

One cluster of annotation approaches, typified by RED or TimeBank (Pustejovsky, Castano et al., 2003; O'Gorman et al., 2016), is one in which annotation is applied by experts over an entire document by selectively labeling informative event–event relations. Such an approach to annotation is naturally compatible with capturing long-distance relations, because the annotators are attempting to craft a representation of the entire timeline using those relations, but limits the density of annotation.

Those approaches differ from the more recent trend of annotation in which corpora are annotated with dense, and often decompositional, temporal relations annotations – providing labels for all possible event–event relations, modulo some set of constraints (Chambers et al., 2014; Ning et al., 2018; Vashishtha et al., 2019).

This characterization into those two prototypical approaches to temporal annotation also allows us to highlight the exceptions to this characterization.

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

Table 7.4. *Summarization of corpora*

Prototypically dense/decompositional	Vashishtha et al. (2019); Chambers et al. (2014); Ning et al. (2018)
Edge cases	Short texts (Mostafazadeh et al., 2016) unconstrained (Hong et al., 2016; Minard et al., 2016), temp dependencies (Kolomiyets et al., 2012; Zhang and Xue, 2018), storylines-focused (Caselli and Vossen, 2017)
Prototypically hierarchical/expert-annotated	Pustejovsky, Hanks et al. (2003); O'Gorman et al. (2016); Styler et al. (2014); Glavaš et al. (2014)

For example, whereas individually examining all possible event pairs tends to naturally require constraints to only consider event-pairs in neighboring sentences, approaches that filter by event ontologies can examine all event–event pairs (Hong et al., 2016) because that filtering reduces the total number of events. Other exceptions are those in which there are other constraints on which relations to annotate, such as constraining relations to match temporal dependencies (Kolomiyets et al., 2012; Zhang and Xue, 2019) or narrative structure (Caselli and Vossen, 2017).

This characterization of these different annotations, summarized in Table 7.4, shows that an important ongoing question for the future of event–event relations corpora is to understand the actual impact of these annotation decisions. We therefore highlight relations that are captured by RED but not currently annotated in any of the dense, decompositional approaches and propose methods for measuring the impact of such annotation decisions in terms of their contribution of information about the actual timeline of events.

7.2.7 Corpus Size Comparisons

We note that many of these corpora provide meaningfully large sets of annotated data, but none are overwhelmingly large. Table 7.5 illustrates the general size of the RED released corpus and of an additional corpus of RED data annotated for the DARPA Active Interpretation of Disparate Alternatives (AIDA) program. Though annotations are often measured in terms of the number of relations annotated, we might count or omit the presence of temporal links to the document time – a simple but informative feature for determining a timeline. Counts including those links are included parenthetically in the last column. Such counts reveal the importance of generating connections

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

Table 7.5. *Size of RED and other temporal corpora; relations include causal and plot relations*

	Events	Times	Event Clusters	Tokens	Relations (w/ DocTime)
RED (2016 release)	8,731	1,127	2,390	54,287	4,969 (13,700)
RED-AIDA	3,639	354	1,801	21,438	1,735 (5,374)
all RED	12,370	1,481	4,191	75,725	6,704 (19,074)
TimeBank (v1.2)	7,935	1,414		61,418	6,418
ECB (v0.9)	7,275	1,297	7,671		(12,423)
TB-Dense	1,729	289			(12,715)
UDS-T	32,302			254,830	70,368
Hong et al. (2016)	863				25,610
Ning et al. (2019)	1,300				3,572

ECB = Event Coreference Bank; TB-Dense = TimeBank Dense; UDS-T = Universal Decompositional Semantics–Temporal.

between annotations such as RED and the larger dense corpora such as UDS-T (Vashishtha et al., 2019) or TimeBank-Dense (Chambers et al., 2014).

7.3 Long-Distance Relations in RED: Contains, Causality, and Coreference

As noted in Sections 7.2.2 and 7.2.4, there are three ways in which temporal information between events separated by a long distance might be marked in RED annotations:

1. Two events are linked by long-distance containment relations (CONTAINS and CONTAINS/SUBEVENT).
2. Two events are linked by a causal relation (because causal relations also capture temporal order).
3. Two events are linked through event coreference, thus allowing temporal transitive closure to infer relations between events linked to those events.

This first focus on containment follows prior discussions of narrative container approaches (Pustejovsky and Stubbs, 2011; Styler et al., 2014), which emphasize the use of CONTAINS (and CONTAINS-SUBEVENT) relations, asking annotators to link an event to the most specific containing span whenever possible. Such larger containing events may have subevents scattered

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

across large distances in some situations, such as when events described in the headline have subevents throughout the article (as in example (4)) or where specific events are part of a larger situation that is unfolding (as in example (5)):

(4) **PROTESTS** AS CITY COLLEGE CLOSES A STUDENT CENTER … (10 sentences). …A protester was **arrested**. [PROTESTS CONTAINS-SUBEVENT arrested]

(5) **FOCUS** TURNS TO THREE SUSPECTS IN BELGIAN DIAMOND ROBBERY … (5 sentences). … Investigators said Bertoldi's arrival in Geneva after the theft had provided a critical breakthrough in the **inquiry**. [Inquiry CONTAINS-SUBEVENT focus]

Similarly, causal annotations are annotated at any distance. Example (6) illustrates one such longer-distance causal link, wherein the event of leaving office is preceded by, and caused by, the later mentions "beat" event, itself a subevent of the 2010 elections:

(6) Skelton …served as the chairman of House Armed Services Committee from 2007 until **leaving** office. … (6 sentences) … Rep. Vicky Hartzler, the Republican who **beat** Skelton and still holds the seat, received support from many Tea Part [beat BEFORE/CAUSE leaving]

Such causal links can be inferred by annotators over long distances. Though allowing annotators to mark any kind of temporal ordering relations over a long distance of text might entail a vast collection of possible temporal order relations, the set of causal relations is much more limited in a text and thus provides more tractable long-distance relations.

7.4 Studying RED Impact on Event Ordering

We do not want to simply make generalizations about which kinds of relations may carry meaningful information without attempting to quantify that impact. However, such a measure is complicated for temporal annotations, because there is no single agreed-upon target of annotation. Instead, temporal annotations provide statements about the partial order of events and might only be truly evaluated against a true timeline of events. Thus, we suggest one way of approaching the evaluation of temporal annotation is to do just that: given a rich way of representing a total order of events in a timeline and a way of estimating that total order based on a given annotation, we can simply compare the correlation between the two (in this case, using Spearman's rank-order correlation).

We present one simple route toward a richer representation of the timeline, relying upon the observation that annotations such as RED are somewhat orthogonal to annotations such as TimeBank-Dense or UDS-T in which an emphasis is placed on dense, exhaustive annotation between adjacent sentences. Merging the contributions of different annotation schemes can naturally provide more information than would be practical to annotate in a single annotation pass and provides some separation between the assumptions of any one annotation and how you examine its results. Such an approach has some similarities to generation of timelines from TimeML outputs (see Chapter 4).

As a pilot of this, we implement RED-style event annotation on four documents in the development set of the English Web Treebank (Bies et al., 2012), the corpus also used for UDS-T annotations (Vashishtha et al., 2019). Because the UDS-T annotations take a dramatically different strategy toward the annotation of temporal information (as discussed in Section 7.2.2), we suggest that the union of the two annotations can be highly informative regarding the entire temporal structure of the document. UDS-T annotations are a crowdsourced annotation of event–event pairs within the same sentence or adjacent sentences, where annotators use two sliding scales to mark the relationships between the event spans, allowing a much more nuanced and continuous annotation than that illustrated by discrete labeling of the traditional temporal relationships. Though that annotation also involved event duration annotation, we do not leverage it here; it would naturally be relevant for extensions of this work.

We merge the RED annotations with the UDS-T annotations by aligning events according to their spans and then adding event–event relations to the graph until none are left, converting temporal relations into statements about order relations between the start and end points of each span. We treat this as a somewhat random process – because assertions about how events are ordered may contradict each other, we randomly select the order of relations to add and do not add relations that contradict existing information. We repeat this process many times for a given set of temporal information, so that we get a set of N timelines that are mostly compatible with the annotation provided. We generated such a temporal graph 50 times for each condition and kept the 10 event orders with the largest transitive closure.

Such an approach provides a more formal way to examine the idea motivating annotations such as RED of focusing on "highly informative" relations such as containment and allows one to measure how much redundancy exists between different annotation approaches. We compare the timelines generated by the union of RED and UDS-T documents to those produced when one

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

Table 7.6. *Correlation between timelines sampled using all data sources and timelines after certain annotations are removed*

Which timelines to compare	File1	File2	File3	File4	Avg. Δ
All vs All	0.211	0.292	0.185	0.261	
Remove RED DocTimeRel	0.223	0.193	0.167	0.222	−0.036
Remove Coref/Causation/Contains	0.228	0.289	0.162	0.229	−0.010
Remove Coref	0.247	0.283	0.171	0.248	0
Remove Causation	0.226	0.296	0.152	0.242	−0.008
Remove Contains	0.202	0.262	0.191	0.220	−0.019
RED only	0.271	0.177	0.137	0.228	−0.034
UDS-T only	0.146	0.199	0.115	0.209	−0.070

RED = Richer Event Description; UDS-T = Universal Decompositional Semantics–Temporal.

omits various pieces of information to study the impact of those individual annotations on the overall understanding of a document timeline. Table 7.6 illustrates, for each such condition, average Spearman's rank-order correlation between the orders (because we generate 10 timelines for each, this is an average of 100 comparisons).

The results of Table 7.6 illustrate the highly variable nature of this approach (most notable that even the correlation between the same merged set of annotations is often quite far from the 1.0 you would expect from an identical set of events) but still serves to show in general terms the impact of different annotation components on a final temporal ordering. Most notable, one can see the extent to which the two annotations provide different information about the document and have much more information than either approach alone. Moreover, one can see that the largest impact from RED annotations is the simple use of links to document time, which provide a simple grounding of each event into the correct general part of the timeline and may be especially useful for events with few other temporal relations. One can also see that of the discussed ways of getting long-distance temporal relations, the annotation with the most impact is that of the temporal containment relations.

7.5 Conclusions

The RED formalism remains one of the most comprehensive forms of event–event relation annotations. However, it is one point within an increasingly large

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

landscape of different temporal relations corpora. We illustrate here a deeper focus on RED-style annotations of temporal relationships in text and illustrate the value of such a richly annotated event-centric corpus.

The evaluation results presented here illustrate a very preliminary way of evaluating such corpora. By attaining rich representations of event orderings and measuring the contribution of annotations to that order, we suggest that one may better study which kinds of annotations are actually informative and which annotations may provide little information. These results also highlight which aspects of RED annotation go beyond what is captured by many other corpora in which dense, local annotations are used. These aspects, such as long-distance temporal containment relations or relationships to speech time, point to important phenomena that researchers building future temporal annotations might consider.

References

Allen, James. 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, **26**(11), 832–843.

Araki, Jun, Liu, Zhengzhong, Hovy, Eduard, and Mitamura, Teruko. 2014. Detecting Subevent Structure for Event Coreference Resolution. Pages 4553–4558 of: Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association.

Araki, Jun, Mulaffer, Lamana, Pandian, Arun, et al. 2018. Interoperable Annotation of Events and Event Relations across Domains. Pages 10–20 of: Bunt, Harry (ed.), *Proceedings 14th Joint ACL–ISO Workshop on Interoperable Semantic Annotation*. Santa Fe, NM: Association for Computational Linguistics.

Bejan, Cosmin Adrian, and Harabagiu, Sanda. 2010. Unsupervised Event Coreference Resolution with Rich Linguistic Features. Pages 1412–1422 of: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Bies, Ann, Mott, Justin, Warner, Colin, and Kulick, Seth. 2012. English Web Treebank. Philadelphia: Linguistic Data Consortium.

Caselli, Tommaso, and Vossen, Piek. 2017. The Event StoryLine Corpus: A New Benchmark for Causal and Temporal Relation Extraction. Pages 77–86 of: Caselli, Tommaso, Miller, Ben, van Erp, Marieke, et al. (eds.), *Proceedings of the Events and Stories in the News Workshop*. Vancouver: Association for Computational Linguistics.

Chambers, Nathanael, Cassidy, Taylor, McDowell, Bill, and Bethard, Steven. 2014. Dense Event Ordering with a Multi-pass Architecture. *Transactions of the Association for Computational Linguistics*, **2**, 273–284.

Croft, William, Pešková, Pavlína, and Regan, Michael. 2016. Annotation of Causal and Aspectual Structure of Events in RED: A Preliminary Report. Pages 8–17

https://doi.org/10.1017/9781108854221.010 Published online by Cambridge University Press

of: Palmer, Martha, Hovy, Ed, Mitamura, Teruko, and O'Gorman, Tim (eds.), *Proceedings of the Fourth Workshop on Events*. Austin, TX: Association for Computational Linguistics.

Cybulska, Agata, and Vossen, Piek. 2014. Using a Sledgehammer to Crack a Nut? Lexical Diversity and Event Coreference Resolution. Pages 4545–4552 of: Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, et al. (eds.), *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*.

Glavaš, Goran, Šnajder, Jan, Kordjamshidi, Parisa, and Moens, Marie-Francine. 2014. HiEve: A corpus for Extracting Event Hierarchies from News Stories. Pages 3678–3683 of: Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, et al. (eds.), *Proceedings of 9th Language Resources and Evaluation Conference*. European Language Resources Association.

Govindarajan, Venkata Subrahmanyan, Van Durme, Benjamin, and White, Aaron Steven. 2019. Decomposing Generalization: Models of Generic, Habitual, and Episodic Statements. *Transactions of the Association for Computational Linguistics*, **7**, 501–517.

Hong, Yu, Zhang, Tongtao, Horowit-Hendler, Sharone, Ji, Heng, O'Gorman, Tim, and Palmer, Martha. 2016. Building a Cross-document Event–Event Relation Corpus. Pages 1–6 of: Friedrich, Annemarie, and Tomanek, Katrin (eds.), *Proceedings of LAW X – The 10th Linguistic Annotation Workshop*.

Ikuta, Rei, Styler, Will, Hamang, Mariah, O'Gorman, Tim, and Palmer, Martha. 2014. Challenges of Adding Causation to Richer Event Descriptions. Pages 12–20 of: Mitamura, Teruko, Hovy, Ed, and Palmer, Martha (eds.), *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. Baltimore, MD: Association for Computational Linguistics.

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

Lee, Heeyoung, Recasens, Marta, Chang, Angel, Surdeanu, Mihai, and Jurafsky, Dan. 2012. Joint Entity and Event Coreference Resolution across Documents. Pages 489–500 of: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*.

Minard, Anne-Lyse, Speranza, Manuela, Urizar, Ruben, et al. 2016. MEANTIME, the NewsReader Multilingual Event and Time Corpus. Pages 4417–4422 of: Calzolari, Nicoletta, Choukri, Khalid, Declerck, Thierry, et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association.

Mirza, Paramita, Sprugnoli, Rachele, Tonelli, Sara, and Speranza, Manuela. 2014. Annotating Causality in the TempEval-3 Corpus. Pages 10–19 of: Kolomiyets, Oleksandr, Moens, Marie-Francine, Palmer, Martha, Pustejovsky, James, and Bethard, Steven (eds.), *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*. Gothenburg, Sweden: Association for Computational Linguistics.

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

Semantic Annotation of Event Structures. Pages 51–61 of: Palmer, Martha, Hovy, Ed, Mitamura, Teruko, and O'Gorman, Tim (eds.), *Proceedings of the Fourth Workshop on Events*. San Diego: Association for Computational Linguistics.

Ning, Qiang, He, Hangfeng, Fan, Chuchu, and Roth, Dan. 2019. Partial or Complete, That's the Question. Pages 2190–2200 of: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

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

O'Gorman, Tim, Wright-Bettner, Kristin, and Palmer, Martha. 2016. Richer Event Description: Integrating Event Coreference with Temporal, Causal and Bridging Annotation. Pages 47–56 of: Caselli, Tommaso, Miller, Ben, van Erp, Marieke, Vossen, Piek, and Caswell, David (eds.), *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*. Austin, TX: Association for Computational Linguistics.

Pustejovsky, James, and Stubbs, Amber. 2011. Increasing Informativeness in Temporal Annotation. Pages 152–160 of: Ide, Nancy, Meyers, Adam, Pradhan, Sameer, and Tomanek, Katrin (eds.), *Proceedings of the 5th Linguistic Annotation Workshop*. Portland, OR: Association for Computational Linguistics.

Pustejovsky, James, Hanks, Patrick, Saurí, Roser, et al. 2003. The TimeBank Corpus. Pages 647–656 of: Archer, Dawn, Rayson, Paul, Wilson, Andrew, and McEnery, Tony (eds.), *Proceedings of Corpus Linguistics Conference*. Lancaster, UK.

Pustejovsky, James, Castano, José, Ingria, Robert, et al. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. AAAI Technical Report SS-03-07.

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

Rudinger, Rachel, White, Aaron Steven, and Van Durme, Benjamin. 2018. Neural Models of Factuality. Pages 731–744 of: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Volume 1. Long Papers*. New Orleans, LA: Association for Computational Linguistics.

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

Song, Zhiyi, Bies, Ann, Strassel, Stephanie, et al. 2015. From Light to Rich ERE: Annotation of Entities, Relations, and Events. Pages 89–98 of: Hovy, Ed, Mitamura, Teruko, and Palmer, Martha (eds.), *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*.

Styler, William F., IV, Bethard, Steven, Finan, Sean, et al. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, **2**, 143–154.

Vashishtha, Siddharth, Van Durme, Benjamin, and White, Aaron Steven. 2019. Fine-Grained Temporal Relation Extraction. Pages 2906–2919 of: *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics.

Vigus, Meagan, Van Gysel, Jens E. L., and Croft, William. 2019. A Dependency Structure Annotation for Modality. Pages 182–198 of: Xue, Nianwen, Croft, William, Hajic, Jan, et al. (eds.), *Proceedings of the First International Workshop on Designing Meaning Representations*. Florence, Italy: Association for Computational Linguistics.

Zhang, Yuchen, and Xue, Nianwen. 2018. Structured Interpretation of Temporal Relations. In: Calzolari, Nicoletta, Choukri, Khalid, Cieri, Christopher, et al. (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: European Languages Resources Association.

Zhang, Yuchen, and Xue, Nianwen. 2019. Acquiring Structured Temporal Representation via Crowdsourcing: A Feasibility Study. Pages 178–185 of: Mihalcea, Rada, Shutova, Ekaterina, Ku, Lun-Wei, Evang, Kilian, and Poria, Soujanya (eds.), *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*.