Interoperable Annotation of Events and Event Relations across Domains

Jun Araki*, **Lamana Mulaffer**†, **Arun Pandian**†,
Yukari Yamakawa*, **Kemal Oflazer**† **and Teruko Mitamura***
*Carnegie Mellon University, Pittsburgh, PA 15213, USA
†Carnegie Mellon University in Qatar, PO Box 24866, Doha, Qatar
junaraki@cs.cmu.edu, fmulaffe@andrew.cmu.edu, apandian@andrew.cmu.edu,
yukariy@andrew.cmu.edu, ko@cs.cmu.edu, teruko@andrew.cmu.edu

Abstract

This paper presents methodologies for interoperable annotation of events and event relations across different domains, based on notions proposed in prior work. In addition to the interoperability, our annotation scheme supports a wide coverage of events and event relations. We employ the methodologies to annotate events and event relations on Simple Wikipedia articles in 10 different domains. Our analysis demonstrates that the methodologies can allow us to annotate events and event relations in a principled manner against the wide variety of domains. Despite our relatively wide and flexible annotation of events, we achieve high inter-annotator agreement on event annotation. As for event relations, we obtain reasonable inter-annotator agreement. We also provide an analysis of issues on annotation of events and event relations that could lead to annotators' disagreement.

1 Introduction

Events are a key semantic component integral to information extraction and natural language understanding. They are a ubiquitous linguistic phenomenon, appearing in numerous domains, and compose rich discourse structures via various relations between events, forming a coherent story over multiple sentences. However, these properties of events have received relatively little attention in the literature. From the perspective of information extraction, much previous work on events pays attention to domain-specific clause-level argument structure (e.g., attackers kill victims, plaintiffs sue defendants, etc), putting less emphasis on what semantically constitutes events. The formalization focusing on domain-specific clause-level argument structure often involves its own definition of events based on instantiation of event ontology for a particular domain, aimed at automatic extraction of closed-domain events, as illustrated in ACE (Doddington et al., 2004), TAC KBP (Mitamura et al., 2017), PASBio (Wattarujeekrit et al., 2004), and BioNLP (Kim et al., 2009). For clarification, we use the term 'domain' to refer to a specific genre of text, such as biology, finance, and so forth. The closed-domain formalization might be of practical use in some domain-specific scenarios. However, it designs event definitions and annotation schemes arbitrarily within the respective domains. For example, ACE considers resulting states (resultatives)[1] as events, but others might exclude them or include a broader notion of states as events. Therefore, it is questionable whether a collection of the existing heterogenous annotation schemes for closed-domain events adequately contribute to interoperable and consistent annotation of events across domains.

On the other hand, prior work on open-domain events have some limitations with respect to coverage of events and their relations. Lexical databases such as WordNet (Miller et al., 1990), FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005) can be viewed as a superset of event lexicon, and their subtaxonomies seem to provide an extensional definition of events. However, these databases have a narrow coverage of events because they generally do not cover current terminology and proper nouns due to their dictionary nature. For instance, none of WordNet, FrameNet and PropBank cover the proper noun 'Hurricane Katrina'. In addition, they do not provide any principles or guidelines about how to annotate events in text by themselves due to their different focus. TimeML (Pustejovsky et al., 2003) focuses on temporal aspects of events and does not deal with multi-word and generic events. ISO-TimeML (ISO,

[1] One example of resulative events in ACE is "They have been married for 3 years."

2012) provides a wider coverage of events than TimeML, but still has a focus on temporal relations. Event annotation in OntoNotes (Weischedel et al., 2011) is restricted to a small number of event nouns. Also importantly, some of the work mentioned above involves theoretical formalization of some relations between events (e.g., temporal relations), but none of the work provides annotation of a variety of event relations such as event coreference, subevents and causality. Richer Event Description (RED) (Palmer et al., 2016) defines events and their relations in a general manner, but its annotation was performed only in the clinical domain (O'Gorman et al., 2016).

In this work, we present our methodologies for annotating events and relations between events in unrestricted domains. The goal of our annotation project is to provide human-annotated data to build a generation generation application to enhance reading comprehension for English as second language (ESL) students, as a continuous effort of (Araki et al., 2016). Using the notion of eventualities (Bach, 1986) and event nuggets (Mitamura et al., 2015), our event annotation scheme defines events and annotates event spans of text while not assigning any specific event types to them. In that sense, our event annotation is span-oriented, as compared to the traditional argument-oriented annotation of events. As for relations between events, we choose to annotate five relations from the perspective of the goal: event coreference, subevent, causality, event sequence ('after' relations), and simultaneity. To our knowledge, this is the first work that performs human annotation of events and the five relations in unrestricted domains. We believe that this work contributes not only to the goal of the annotation project but also to an important step toward interoperable annotation of events and their relations across domains. The five event relations cover various semantic and temporal aspects of events, deeply connected with common-sense and domain-specific knowledge. Thus, we assume that a mixture of the five relations forms meaningful event structures as semantic backbones to facilitate natural language understanding and sophisticated document-level reasoning. Such event structures are valuable for generating high-level questions for reading comprehension, such as the one that requires learners to infer answers over multiple sentences (Araki et al., 2016).

Our contribution is twofold. First, we annotate a wide coverage of events, comprising verbs, nouns, adjectives, and phrases which are continuous or discontinuous (see Section 3). Despite this relatively wide and flexible annotation of events on text in 10 different domains, we show that our annotation achieved high inter-annotator agreement. Second, unlike previous methodologies which generally focus on deal only with event coreference such as ECB+ (Cybulska and Vossen, 2014), we present methodologies to annotate five event relations in unrestricted domains (see Section 4).

2 Data and Annotation Procedures

In this section, we describe our data and annotation procedures. Our annotation target is not restricted in any specific domains. Thus, ideally speaking, our annotation should include all kinds of events in a domain-agnostic manner. However, annotating all kinds of events manually in unrestricted domains would be unrealistic due to annotation cost. Therefore, in order to make the corpus creation manageable while retaining the domain diversity, we select 100 articles in Simple English Wikipedia[2], comprising 10 from each of 10 different domains.[3] The domains are: architecture, chemistry, disasters, diseases, economics, education, geology, history, politics, and transportation. We choose Simple Wikipedia because our annotators are not necessarily experts in these domains, and the simplified sentences could facilitate our annotation of events and event relations against text from the wide variety of domains. We refer to the corpus as **SW100**.

Our annotation is done by two annotators and a more experienced annotator whom we call the adjudicator. We first write our annotation guidelines to guide how to annotate events and event relations. We then set up an initial learning period in which the three annotators learn how to annotate events through answering their questions. We use BRAT (Stenetorp et al., 2012) as an annotation tool. We take a two-stage approach: (1) annotating and finalizing events and (2) annotating and finalizing event relations. Finalization is the adjudicator's process of comparing annotations, resolving annotation differences, and

[2] `https://simple.wikipedia.org`

[3] For more detailed corpus statistics, see Section 3.3 for events and Section 4.3 for event relations.

producing a single set of annotations. More detailed steps are as follows:

1. Three annotators identify event spans, following the annotation guidelines.
2. We compute inter-annotator agreement on event annotation.
3. The adjudicator finalizes event annotation.
4. Three annotators identify event relations on top of the finalized events, following the annotation guidelines.
5. We compute inter-annotator agreement on event relation annotation.
6. The adjudicator finalizes event relation annotation.

3 Annotation of Events

This section describes our definition of events and principles for annotation of events.

3.1 Definition of Events: Eventualities

As with TimeML (Pustejovsky et al., 2003) and ISO-TimeML (ISO, 2012), our definition of events uses *eventualities* (Bach, 1986), which are a broader notion of events, including states, processes, and events. This definition is inclusive in the sense that it includes states in addition to events and processes. We define the three classes on the basis of durativity and telicity (Moens and Steedman, 1988; Pulman, 1997):

- **states**: notions that remain unchanged until their change or are brought as a result of an event, e.g., He **owns** a car. Tom was **happy** when he received a present;
- **processes**: notions that involve a change of state without an explicit goal or completion, e.g., it was **raining** yesterday;
- **events**[4]: notions that involve a change of state with an explicit goal or completion, e.g., **walked** to Boston, **buy** a book.

We recognize that annotating states is generally more difficult than annotating processes and actions because states are often confused with attributes which are not eventive and thus should not be annotated. For example, let us consider the following examples:

(1) Mary was **talkative** at the **party**.
(2) Mary is a *talkative* person.

In (1), 'talkative' is eventive because it implies that Mary talked a lot at the party, whereas 'talkative' in (2) is not because it just indicates Mary's personal attribute. Note that we introduce the notion of eventualities in order to clarify the semantic boundary between eventive and non-eventive, not because we are interested in classifying events into the three classes of actions, processes, and states.

3.2 Annotation of Events: Event Nuggets

We also define what textual units are annotated as events. For this purpose, we use the notion of *event nugget* (Mitamura et al., 2015). An event nugget is defined as a semantically meaningful unit that expresses an event. It can be either a single word (verb, noun, or adjective) or a phrase which is continuous or discontinuous, depending on how we interpret the semantic meaningfulness of an event that the event nugget refers to. We give several examples below, where we use boldface to highlight event nuggets and underlines to show units of multi-word ones.

(3) The gunman **shot** the teller in the bank.
(4) The gunmen **opened fire** at the teller in the bank.
(5) I **cried** when my grandpa **kicked** the **bucket**.
(6) Susan **turned** the TV **on**.
(7) She **responded** his email **dismissively**.

[4]These notions were named 'transitions' in ISO-TimeML.

In (3), 'shot' is the only verb representing an event, and we annotate 'shot' as a single-word event nugget. On the other hand, in (4) we annotate 'open fire' as a single multi-word event nugget because the phrase 'open fire' indicates more complete meaning than either 'opened' or 'fire.' Similarly, in (6) we annotate 'turned ... on' as a single discontinuous multi-word event nugget, excluding 'the TV'. As a result, we can consider the phrase semantically meaningful and annotate it as a single event nugget. In (7), we annotate 'dismissively' as a part of an event nugget because it implies the action of her dismissing.

3.3 Corpus Analysis of Events

We show statistics of event annotations in SW100 in Table 1. Multi-word event nuggets amount to 955. 24% of the 955 are discontinuous, and most (97%) of the discontinuous multi-word event nuggets are verb phrases. 'Others' in Table 1(b) include pronouns, demonstrative determiners, and numbers.[5]

Domain	# (%)	Domain	# (%)
Architecture	475 (8.8)	Education	653 (12.1)
Chemistry	576 (10.7)	Geology	483 (8.9)
Disaster	510 (9.4)	History	486 (9.0)
Disease	618 (11.4)	Politics	534 (10.0)
Economics	479 (8.9)	Transportation	583 (10.8)

(a) Event nuggets with respect to domains.

	Single-word	Multi-word	All
Verb	2799 (51.9)	560 (10.4)	3359 (62.2)
Noun	1273 (23.6)	382 (7.1)	1655 (30.6)
Adjective	192 (3.6)	2 (0.0)	194 (3.6)
Others	178 (3.3)	11 (0.2)	189 (3.5)
All	4442 (82.3)	955 (17.7)	5397 (100.0)

(b) Event nuggets with respect to syntactic types.

Table 1: Statistics of events in SW100. Percentages (%), shown in parentheses, indicate ratios to the total number of event nuggets (i.e., 5397).

3.4 Inter-annotator Agreement on Event Annotation

Event annotation involves annotation of text spans. Thus, we measure inter-annotator agreement using the pairwise F1 score under two conditions: strict match and partial match. The former checks whether two annotations have exactly the same span. The latter checks whether there is an overlap between annotations, with the restriction that each annotation can only be matched to one annotation by the other annotator. Regarding the adjudicator's annotation as gold standard, we compute a pairwise F1 score between the adjudicator and one of the other two annotators, and another pairwise F1 score between the adjudicator and the other annotator. We then take the average of the F1 scores as our inter-annotator agreement. As a result, the inter-annotator agreement was 80.2% (strict match) and 90.2% (partial match).

3.5 Issues on Annotation of Events

The main challenge of event annotation is ambiguities on eventiveness. No matter how well eventiveness is defined, there is a lot of discretion required from the subjective viewpoint of the annotator to resolve ambiguous cases. For instance, let us consider the following sentences:

(8) These were issues of interest like the welfare state.
(9) Force equals mass times acceleration.

It is not entirely clear whether 'issues' in (8) should be annotated as an event nugget as it may constitute a set of events as well as a set of non-events. 'Force' in (9) is associated with its physical sense, but still determining whether it is eventive is difficult. Therefore, we conjecture that deciding eventiveness is not a clear binary classification problem, and there exists a continuum between eventive and non-eventive in the space of event semantics.

Another type of ambiguities arises from semantic meaningfulness of event spans in the definition of event nuggets. When we annotate multi-word event nuggets, it can be unclear which set of words constitute a semantically meaningful unit. For example, let us consider the following sentence:

(10) Bricks are used in masonry construction.

[5]Examples of the pronouns and demonstrative deteriminers are 'it' and 'this event', respectivetly, referring to previously mentioned events. An example of the numbers is 'one', which also refers a previously mentioned event.

One interpretation is that 'masonry construction' in (10) can be seen as a semantically meaningful unit. In contrast, another plausible interpretation is that only 'construction' is a semantically meaningful unit while 'masonry' is considered a mere specifier. Other challenges include knowing idiomatic expressions to annotate event nuggets. Because such expressions are often cultural, it could be an challenge to arrive at an consensus on idiomatically expressed events such as 'kicked the bucket'.

4 Annotation of Event Relations

In this section, we define event relations that we annotate and describe our principles of annotation of event relations.

4.1 Definition and Annotation Principles of Event Relations

As mentioned in Section 1, we define and annotate 5 event relations: event coreference, subevent, causality, event sequence, and simultaneity.

Event coreference. We define event coreference as a linguistic phenomenon that two event nuggets refer to the same event. For two event nuggets to corefer, they should be semantically identical, have the same participants (e.g., agent, patient) or attribute (e.g., location, time), and have the same polarity. For instance, 'Great Fire of London' and 'fire' are coreferential in (11).

(11) The **Great Fire of London** happened in 1666. The **fire** lasted for three days.

When considering event identity for event coreference, we use the notion of *event hopper* from Rich ERE (Song et al., 2015), which is a more inclusive, less strict notion than the event coreference defined in ACE.

Subevent. Following (Hovy et al., 2013), we define subevent relations as follows. Event A is a subevent of event B if B represents a stereotypical sequence of events, or a script (Schank and Abelson, 1977), and A is a part of that script. For example, 'affected', 'flooded' and 'broke' are three subevents of 'Hurricane Katrina' in (12). We refer to 'Hurricane Katrina' as a parent (event) of the three subevents.

(12) On August 29, 2005, New Orleans was **affected** by **Hurricane Katrina** which **flooded** most of the city when city levees **broke**.

Causality. We define causality to be a cause-and-effect relation, in which we can explain the causation between two event nuggets X and Y, saying "X causes Y". One example of causality is "The **tsunami** was caused by the **earthquake**." Causality also adds another distinctive characteristic to annotation. Causality inherently entails an event sequence. For example, if we say "The **tsunami** was caused by the **earthquake**", it means that the tsunami happened after the earthquake. To distinguish causality from event sequences and other relations such as preconditions (Palmer et al., 2016), we perform causality tests, largely based on (Dunietz et al., 2017):

1. The "why" test: After reading the sentence, can an annotator answer "why" questions about the potential effect argument? If not, it is not causal.
2. The temporal order test: Is the cause asserted to precede the effect? If not, it is not causal.
3. The counterfactuality test: Would the effect have been just as probable to occur or not occur had the cause not happened? If so, it is not causal.
4. The ontological asymmetry test: Could you just as easily claim the cause and effect are reversed? If so, it is not causal.
5. The linguistic test: Can the sentence be rephrased as It is because (of) X that Y or X causes Y? If so, it is likely to be causal.
6. The granularity test: Does X have the same event granularity as Y? If not, it is not causal. We define event granularity to be the scale of an event indicating how large the event is (e.g., events 'lunch' and 'dinner' have the same granularity, but 'ordering' is a finer-grained event than dinner. We add a constraint that two events with causality must have the same event granularity.

Event sequence. We define event sequence ('after' links) as follows. If event A is after event B, A happens after B happens under stereotypicality within a script or over multiple scripts. Note that we

do not consider an event sequence relation just by the chronological order to avoid trivial (nonsensical) sequence relationships. We give an example from a restaurant script:

(13) We **went**(E1) to **dinner**(E2) at a famous restaurant. We **ordered**(E3) steak and **ate**(E4) it. We then **got** a **call**(E5). After the **call**(E6), we **paid**(E7) and **left**(E8) the restaurant.

In this example, we annotate event sequence relations as follows: E1 $\xrightarrow{\text{after}}$ E3 $\xrightarrow{\text{after}}$ E4 $\xrightarrow{\text{after}}$ E7 $\xrightarrow{\text{after}}$ E8. Note that we do not annotate 'after' links between E4 and E5 and between E6 and E7, even an explicit discourse marker 'After' at the beginning of the third sentence. This is because we do not see stereotypicality of the restaurant script in the sequence of E4, E5 and E6. The merit of our script-based approach for event sequences is that we can still sequence events utilizing the streotypicality of a script when temporal information is not explicitly provided in texts. In addition, it also allows us to frame texts in a story-like structure because of the scripts that we identify in the course of annotation, which is suitable to the goal of our annotation project described in Section 1.

Simultaneity. We define simultaneity as a relation that two event nuggets occur at the same time. We add a 'simultaneous' link between events when you realize that those events take place at the same time. The conjunctions such as 'when' and 'while' are clear markers for simultaneity. We give some examples:

(14) My boss was **talking** over the phone when I **stopped by** his office.
(15) **Right-click** on the mouse button while **holding down** the Shift key.

When annotating simultaneity, we make sure that two events connected with a simultaneous link have the same event granularity. This is helpful to differentiate simultaneity from subevent relations. Below are two examples:

(16) He **kept quiet** during the **meeting**.
(17) He **testified** in the **trial**.

In (16), 'kept quiet' and 'meeting' happened at the same time, but we do not annotate a simultaneity relation between them because they do not have the same granularity. Similarly, we do not annotate a simultaneity relation between 'testified' and 'trial' in (17).

Unlike event sequence, we do not consider the notion of scripts in the case of simultaneity. Considering scripts in annotating simultaneity seems rather restrictive. From the goal of the annotation project, we conjecture that we can annotate many useful simultaneity relations by not considering scripts.

4.2 Facilitating Annotation of Event Relations

As described in Section 2, we use the BRAT tool to annotate events and event relations. However, our initial analysis reveals that although the original version[6] of BRAT supports annotation of relations within a single sentence well, it hinders annotation of relations spanning multiple sentences significantly due to its sentence-oriented visualization. In particular, if one annotates many relations over multiple sentences, a stack of the corresponding horizontal arrows make a working screen too vertically long for human annotators to perform annotation. Therefore, we modified BRAT so that events in different sentences can be directly connected by straight arrows without expanding the screen too much. This modification improves BRAT's visualization and facilitates our event relation annotation greatly. Figure 1 illustrates the improved visualization with some examples of our annotation of events and event relations.

4.3 Corpus Analysis of Event Relations

Table 2 shows statistics of the corpus. As for event coreference, we count the number of event coreference clusters instead of the number of event coreference relations. An event coreference cluster means a cluster grouping two or more coreferential events. For the other four event relations, we adopt the notion of link propagation by Mitamura et al. (2017) and only count relations between event coreference clusters, which avoids counting semantically equivalent (redundant) relations. For instance, if we have

[6]This is the latest version v1.3, available at `http://brat.nlplab.org/`.

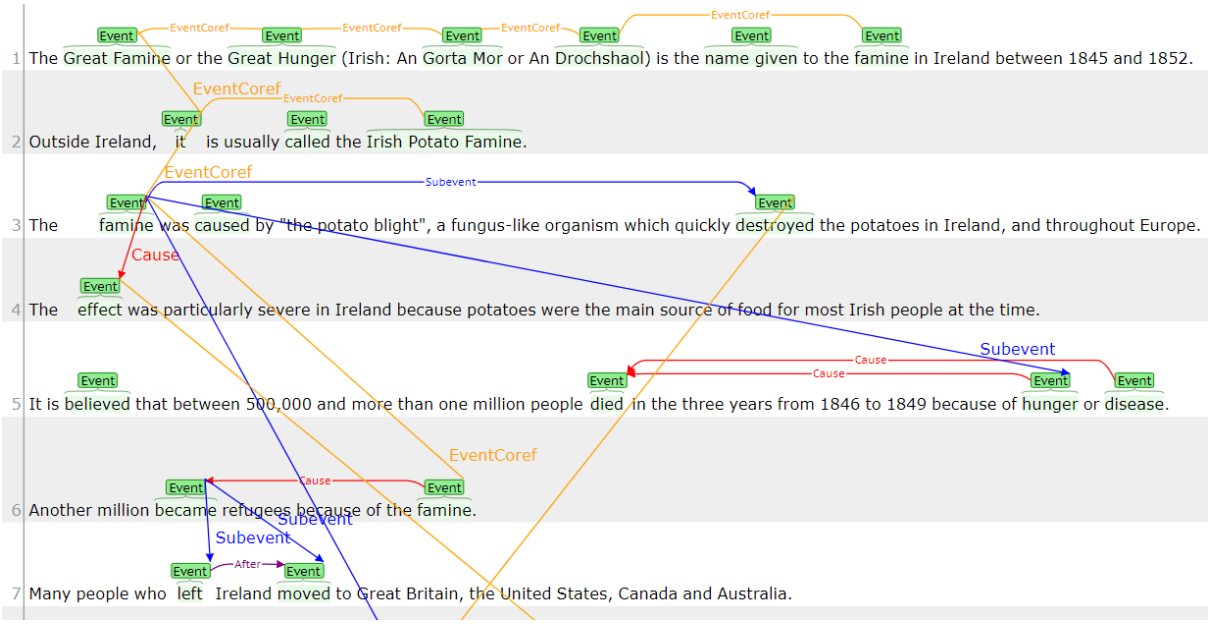

Figure 1: Some examples of our annotation of events and event relations. Our modified version of BRAT facilitates human annotation of event relations adequately, supporting relations spanning multiple sentences along with relations within a single sentence in compact visualization.

event coreference between E1 and E2 and two subevent relations from E1 to E3 and from E2 to E3, we count one event coreference cluster [E1,E2] and one (cluster-level) subevent relation [E1,E2] $\xrightarrow{\text{subevent}}$ [E3]. Overall, subevents are most frequently annotated among the four event relations, and the number of subevent relations is 6.5 times larger than that of simultaneity relations. It is also interesting to see that causality relations appear more frequently in the disease, geology and disaster domains than others. We observed that scientific domains such as diseases and geology generally tend to have more cause-and-effect relationships, but also often require domain-specific knowledge to distinguish between causality and preconditions, as seen in chemical reactions. We also found that the history domain has the largest number of event sequences and the second largest number of subevent relations, indicating that script-based structures tend to appear more frequently in that domain.

	Arch	Chem	Disa	Dise	Eco	Edu	Geo	Hist	Poli	Tran	Total
# Event coreference clusters	32	61	49	67	51	61	52	42	46	51	512
# Subevent relations	74	81	98	94	48	101	91	119	147	93	946
# Causality relations	31	63	71	105	37	28	87	68	36	60	586
# Event sequence relations	48	38	63	30	73	62	58	117	66	88	643
# Simultaneity relations	8	11	21	10	15	17	8	21	24	11	146

Table 2: Statistics of event coreference clusters and cluster-level event relations in SW100. For brevity, we use a prefix with 3 or 4 characters to refer to each domain.

4.4 Inter-annotator Agreement on Annotation of Event Relations

One way to compute inter-annotator agreement on event coreference is to use evaluation metrics developed by prior work, such as MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), and BLANC (Recasens and Hovy, 2011). However, these metrics are appropriate specifically for (event) coreference and cannot be consistently applied to other event relations such as subevents. Instead, a single consistent metric is ideal for comparing inter-annotator agreement. Since we have three annotators, we use Fleiss' Kappa (Fleiss, 1971) to compute inter-annotator agreement on the five event relations annotated by the three annotators. Specifically, we consider all pairwise relations between events and propagate event relations via event coreference, following (Mitamura et al., 2017). Table 3 shows the

result. According to the interpretation of Fleiss' Kappa by Landis and Koch (1977), the inter-annotator agreement on event coreference is substantial agreement, those on subevent and causality relations are fair agreement, and those on after and simultaneity relations are slight agreement.

Relation	κ
Event coreference	0.645
Subevent	0.223
Causality	0.298
Event sequence	0.139
Simultaneity	0.108

Table 3: Inter-annotator agreement (Fleiss' Kappa) on event relations.

4.5 Issues on Annotation of Event Relations

As compared to the high inter-annotator agreement on events described in Section 3.4, our inter-annotator agreement on event relations (especially on the four relations of subevent, causality, event sequence, and simultaneity) is quite low, as shown in Table 3. This low inter-annotator agreement reflects many difficulties that the annotators faced in their event relation annotation. This section describes the difficulties in detail.

Event annotation error. Missing event nuggets or event nuggets with incorrect spans can end up with false annotation of event relations. Note that we do not make any changes to annotated event nuggets during the process of annotating event relations, since we first finalized event nugget annotation before annotating event relations, as described in Section 2. One example is:

(18) Chronic Obstructive Pulmonary Disease (COPD) can **make** breathing gradually **difficult**. **Breathing difficulties** caused by COPD can be compounded by ...

The first event nugget in (18) is 'make ... difficult'. If 'make breathing ... difficult' were annotated instead, it would be coreferent with 'Breathing difficulties'. However, if event nuggets are annotated as shown above, event coreference should not probably be annotated even if it does exist.

Event granularity. When we annotate a subevent relation between event X and Y, we need to figure out a difference in event granularity between X and Y along with a certain script. However, it is sometimes difficult to discern whether X and Y are expressed at different levels in event hierarchy. For example, let us consider the following sentence:

(19) When Mount St. Helens **erupted** in 1980, it **released** 1000 times less material.

The first event 'erupted' can be seen as a parent event of 'released' under the eruption script. Another interpretation is that the two events have the same granularity and there is a causality relation: 'erupted' $\xrightarrow{\text{cause}}$ 'released'. We often need to examine surrounding contexts deeply to resolve the ambiguity.

Script identification. The identification of scripts and their underlying subevents depends largely on common-sense knowledge and intuition of annotators. Thus, it is not easy to arrive at an consensus for subevent relations and event sequences. One example is:

(20) He **sought treatment** for his **cancer**, after which he **got better**.

Some annotators might decide that this sentence constitutes a sickness script which corresponds to the typical life cycle of falling sick and recovering from it. However, this kind of decision can also be subjective, depending on annotators' common-sense knowledge, and others might not admit stereotypicality in the set of events.

Domain-specific knowledge. Annotation of causality and subevent relations can require annotators to have extensive background knowledge. This difficulty can break annotators' agreement easily. We give two examples:

(21) The **start** of the **Cultural Revolution** **followed** the **failure** of the **Great Leap Forward**. Mao **tried** to **remove** capitalists from the Communist Party of China.

(22) The 1973 **oil crisis** **started** on October 17, 1973, when the members of Organization of Arab Petroleum Exporting Countries (OAPEC) **said**, because of the **Yom Kippur War**, that they would no longer **ship** petroleum to nations that had **supported** Israel in its **conflict** with Syria and Egypt.

In (21), whether event 'tried (to remove capitalists)' is a subevent of 'Cultural Revolution' is heavily subjective to annotators' knowledge about the Cultural Revolution. The sentence in (22) has seven events, and the correct annotation of event relations among them requires comprehensive understanding of the 1973 oil crisis and the Yom Kippur War at least, which can cause annotators' disagreement.

Causality vs. Event sequence. Since causality connotes an event sequence by its nature, we employ causality tests to differentiate causality from event sequences, as described in Section 4.1. Even with the causality tests, there are still some ambiguous cases:

(23) Igneous rock can **melt** into magma, **erode** into sediment, or be **pressed** tightly together to **become metamorphic**.

Some annotators may see causality between 'pressed' and 'become metamorphic' whereas others may find it to be an event sequence.

Simultaneity vs. Event sequence. It turns out that our definition and annotation principles on simultaneity described in Section 4.1 are not completely informative with respect to how to deal with the duration of events. For example, it is not entirely clear whether we should annotate simultaneity when two events overlap in time to a large extent but not fully. As a result, this issue confuses annotators when they annotate simultaneity or event sequences. We give two examples:

(24) When 1,500 missiles were **shipped**, three hostages were **released**.

(25) A person can **have dyslexia** even if he or she is very smart or **educated**.

In one interpretation of (24), 'released' happened at the same time as 'shipped'. Another possible interpretation is that 'released' happened after 'shipped'. We observe similar ambiguity in (25).

5 Conclusion and Future Work

We have presented our methodologies for annotating events and five types of event relations: event coreference, subevents, causality, event sequence, and simultaneity. To our knowledge, this is the first work that performs human annotation of events and five event relations of event coreference, subevent, causality, event sequence and simultaneity in a domain-agnostic manner. Using 100 articles in Simple Wikipedia from 10 different domains, we have demonstrated that the methodologies can allow us to annotate a wider coverage of events and event relations than prior work in a principled manner against the wide variety of domains. In addition, we have achieved high inter-annotator agreement on event annotation. Given lower inter-annotator agreement on event relation annotation, we have provided an analysis of issues on annotation of event relations.

There are a number of avenues for future work. The main piece of future work is to improve inter-annotator agreement on event relations by refining the annotation principles and guidelines. It is necessary to develop a more sophisticated annotation scheme for differentiating between subevent relations, causality, event sequences, and simultaneity. As for more temporarily-oriented relations, such as event sequences and simultaneity, we need to introduce a consistent principle for annotators to comprehend the duration of events more precisely, such as interval temporal logic (Allen, 1983). Besides the five event relations that we dealt with in this work, there exist many other event relations such as memberships (Hovy et al., 2013) and bridging (Palmer et al., 2016). Providing more comprehensive annotation guidelines including these event relations would also lead to an improvement on inter-annotator agreement. With respect to event annotation, one could define event spans more adequately than the 'semantically meaning unit' in the event nugget definition, thereby reducing annotators' subjective discretion and improving inter-annotator agreement.

Acknowledgements

This publication was partly made possible by grant NPRP-08-1337-1-243 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING*, pages 1125–1136.

Emmon Bach. 1986. The algebra of events. *Linguistics and Philosophy*, 9:5–16.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC Workshop on Linguistics Coreference*, pages 563–566.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING*, pages 86–90.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of LREC*, pages 4545–4552.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program tasks, data, and evaluation. In *Proceedings of LREC*, pages 837–840.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Proceedings of NAACL-HLT Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28.

ISO. 2012. ISO 24617-1 language resource management – semantic annotation framework – part 1: Time and events. In *International Standardization Organization*.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of BioNLP-ST Workshop*, pages 1–9.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT/EMNLP*, pages 25–32.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event nugget annotation: Processes and issues. In *Proceedings of NAACL-HLT Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 66–76.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2017. Events detection, coreference and sequencing: What's next? Overview of the TAC KBP 2017 Event track. In *Proceedings of Text Analysis Conference*.

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

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*, pages 47–56.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Martha Palmer, Will Styler, Kevin Crooks, and Tim O'Gorman, 2016. *Richer Event Description (RED) Annotation Guidelines*. University of Colorado at Boulder. Version 1.7, `https://github.com/timjogorman/RicherEventDescription/blob/master/guidelines.md`.

Stephen G. Pulman. 1997. Aspectual shift as type coercion. *Transactions of the Philological Society*, 95(2):279–317.

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, and Graham Katz. 2003. TimeML: Robust specification of event and temporal expressions in text. In *Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 28–34.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From Light to Rich ERE: Annotation of entities, relations, and events. In *Proceedings of NAACL-HLT Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 89–98.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: A Web-based tool for NLP-assisted text annotation. In *Proceedings of EACL: Demonstrations Session*, pages 102–107.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of MUC-6*, pages 45–52.

Tuangthong Wattarujeekrit, Parantu K. Shah, and Nigel Collier. 2004. PASBio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5:155.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 54–63. Springer-Verlag New York.