Event Nugget Annotation: Processes and Issues

Teruko Mitamura, Yukari Yamakawa, Susan Holm
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA
teruko@cs.cmu.edu, {yukariy, sh4s}@andrew.cmu.edu

Zhiyi Song, Ann Bies, Seth Kulick, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania, Philadelphia PA
{zhiyi, bies, skulick, strassel}@ldc.upenn.edu

Abstract

This paper describes the processes and issues of annotating event nuggets based on *DEFT ERE Annotation Guidelines v1.3* and *TAC KBP Event Detection Annotation Guidelines 1.7.* Using Brat Rapid Annotation Tool (brat), newswire and discussion forum documents were annotated. One of the challenges arising from human annotation of documents is annotators' disagreement about the way of tagging events. We propose using Event Nuggets to help meet the definitions of the specific type/subtypes which are part of this project. We present case studies of several examples of event annotation issues, including discontinuous multi-word events representing single events. Annotation statistics and consistency analysis is provided to characterize the inter-annotator agreement, considering single term events and multi-word events which are both continuous and discontinuous. Consistency analysis is conducted using a scorer to compare first pass annotated files against adjudicated files.

1 Introduction

Annotating event mentions is useful for event detection tasks. It also is useful for detecting event coreference, subevent relations, event arguments, and realis values in corpora. This paper describes the processes and issues of annotating event nuggets based on the *DEFT ERE Annotation Guidelines v1.3* (LDC, 2014) (henceforth referred to as *Light ERE Guidelines*) and the *TAC KBP Event Detection Annotation Guidelines v1.7* (LTI, 2014) (henceforth referred to as *TAC KBP Event Guidelines*). Using the Brat Rapid Annotation Tool (brat)[1], we annotated files in newswire and discussion forums genres to create the corpus that supports the TAC KBP pilot evaluation for Event Nugget Detection as part of the DEFT program.

In this paper, we introduce the notion of event nugget and how event nuggets are annotated in the corpus. We discuss the issues that arose in the process of developing *TAC KBP Event Guidelines*, because they are important challenges for manual annotation and impact the quality of annotation for gold standard creation. Two major issues are (1) determining if an event meets the event type/subtype definitions and (2) deciding which words should be tagged within the span of a multi-word event nugget that represents a single event. We provide screen images of our annotation tool in order to give a complete picture of the annotation process. Finally, we present statistics to explain the characteristics of the corpus, such as the size of the corpus and the distribution of event type/subtypes. We discuss consistency analysis of inter-annotator agreement in terms of single word, multi-word continuous, and multi-word discontinuous event nuggets.

2 What is an Event Nugget?

It is challenging to provide clear-cut definitions of events, because many researchers define events differently. For example, in the Light ERE annotations, as well as in ACE*, Automatic Content Extraction) English Annotation Guidelines for Events* (LDC, 2005), an event is defined as an explicit occurrence involving participants. An event is something that happens at a particular place and time, and it can frequently be described as a change of state. The *Light ERE Guidelines* expect annotators to tag an event trigger, which is the smallest extent of text that expresses the occurrence of an event. Both ACE and Light ERE, only examples of a particular set of types/subtypes are tagged. An event trigger is usually a word or phrase. In many cases, event triggers are main verbs in sentences that in-

[1] Brat Rapid Annotation Tool (brat) was developed by Pontus Stenetorp et al. (2014). It is a web-based annotation tool.

Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT 2015, pages 66–76,
Denver, Colorado, June 4, 2015.

dicate the occurrence of the events. Annotating a main verb is relatively easy and is likely to produce a higher rate of inter-annotator agreement, because it allows annotators to pay more attention to a syntactic attribute of an event as well as its semantic feature. However, event triggers are not just verbs. Some nouns and adjectives can also express events (See examples in Section 3.1.).

In this study, we took a different approach to event annotations so that we would be able to annotate more complex events, which consist of multiple words taggable as events. For this reason, we decided to take a semantically oriented approach for annotation. New annotation guidelines were produced (*TAC KBP Event Guidelines*), based on the *Light ERE Guidelines* and *ACE*. To clarify the tagging of multiword events, we propose the idea of "event nugget," which is comprised of a semantically meaningful unit that expresses the event in a sentence. An event nugget can be either a single word (main verb, noun, adjective, adverb) or a continuous or discontinuous multi-word phrase.

The main reason why we propose event nugget annotation is to identify events accurately enough to meet the definitions of event types/subtypes in the *Light ERE Guidelines*. The type/subtype definitions restrict annotation to very specific types of events. Figuring out which events fall within the type/subtype definitions is a key issue to annotation. In the process of annotation, we have encountered cases in which multiple words could equally be considered as an event trigger. In many cases the multiple words are hard to separate from one another in terms of meaning (e.g., "hold a meeting", "serve a sentence", "send email"). Thus, we decided to annotate the maximum extent of text which meets the definition of the event types/subtypes provided by the *Light ERE Guidelines*. This approach allows annotators to tag all possible words that meet the definition of the event types/subtypes.

In addition to the annotation of the maximum extent of events, discontinuous tagging is another characteristic of event nugget annotation. (In order to clarify which words are in the same event nugget in this paper, we underline from the first word in a discontinuous multiword event nugget to the last word in the nugget. A dotted underline appears under words that are not part of the event nugget.) Discontinuous tagging allows annotators to tag words that do not lie next to each other but still belong to a multiword event nugget because they are all required to meet the definition, such as "The company **laid** 10 workers **off**," and "His **death sentence** was **carried out**."

Discontinuous tagging is very effective because it can be used to prevent violations of rules for annotation. For example, *TAC KBP Event Guidelines* as well as *Light ERE Guidelines* mention that non-main verbs should not be tagged. In sentences such as "His death sentence was carried out," annotators may want to tag "death sentence was carried out" to meet the definition of Justice_Execute events, since carrying out a death sentence means executing someone. However, tagging "was" violates the rule that non-main verbs are not taggable. In this case, tagging "death sentence" and "carried out" together as a discontinuous multiword event nugget not only meets the definition of Justice_Execute events but also does not violate the rule that "be" verbs should not be tagged.

The merits of event nugget annotation are summarized as follows: identification of events in a more semantically meaningful way and flexible annotation without violating annotation rules. In the next section, we present examples of event nuggets, using the following format to indicate the annotation: [Event Type_Subtype, REALIS]. Realis will be discussed in Section 3.3.

3 Types of Event Nuggets and REALIS

3.1 Single-Word Event Nuggets

As in ACE and Light ERE annotation, single-word event nuggets meet the definitions of event triggers for particular types/subtypes. Slightly modified in *TAC KBP Event Guidelines*, single-word event nuggets refer to words that meet the definitions of event types/subtypes by themselves. They are verbs (usually main verbs), nouns, adjectives, or adverbs. Below are some examples of single-word event nuggets. The words in **bold face** are event nuggets.

- The **attack** by insurgents occurred on Saturday. [Conflict_Attack, ACTUAL]
- Hillary Clinton was not **elected** president in 2008. [Personnel_Elect, OTHER]

There are some cases where multiple single-word event nuggets appear in the same sentence.

- Kennedy was **shot dead** by Oswald. [Conflict_Attack, ACTUAL], [Life_Die, ACTUAL]
- Three years ago, investors **bought** two stagnant web-hosting companies and **merged** them into what is now known as The Planet. [Transaction_Transfer-Ownership, ACTUAL], [Business_Merge-Org, ACTUAL]

Pronouns and other anaphors are also considered as single-word event nuggets if they refer to previous event mentions that meet the definitions of event types/subtypes.

- The **talks** between the Koreas were largely unsuccessful. **They** ended without agreement on Monday. [Contact_Communicate, ACTUAL], [Contact_Communicate, ACTUAL]

3.2 Complex (Multi-Word) Event Nuggets

Complex event nuggets are multi-word phrases (or compounds) that construct semantic units that meet the definitions of event types/subtypes. Those units can be continuous or discontinuous. Multi-word event nuggets take various forms such as verb+noun, verb+particle/adverb, noun+noun, and so on. The words underlined and in **bold face** are multi-word event nuggets that represent a single event.

- Foo Company had **filed Chapter 11** in 2000. [Business_Declare-Bankruptcy, ACTUAL]
- The police investigated the **murder incident**. [Conflict_Attack, ACTUAL]

Discontinuous tagging is one of the characteristics of annotation of multi-word event nuggets. This type of tagging is useful because it captures event nuggets accurately without missing important components of meaning. Below are the examples of discontinuous tagging of multi-word event nuggets.

- The court **found** him **guilty**. [Justice_Convict, ACTUAL]
- His **death sentence** was **carried out**. [Justice_Execute, ACTUAL]
- All **charges** were **dropped** against him last year. [Justice_Acquit, ACTUAL]

Multi-word event nuggets that represent single events are tagged either continuously or discontinuously depending on the particular construction of the semantic units that meet the definitions of the event types/subtypes in each sentence.

For example, consider the definition of Justice_Sue: "A SUE event occurs whenever a court proceeding has been initiated for the purposes of determining the liability of a PERSON, ORGANIZATION or GPE accused of committing a crime or neglecting a commitment." The three examples below illustrate event nuggets for Justice_Sue events. (For clarification, strikethrough denotes an event that is not part of the event nugget being illustrated.)

- His lawyer should **file** a **lawsuit**. [Justice_Sue, OTHER]
- His lawyer should **sue**. [Justice_Sue, OTHER]
- His lawyer should ~~contest~~ the **lawsuit**. [Justice_Sue, OTHER]

The noun+verb combination of "file" and "lawsuit" meet the definition of Justice_Sue as a court proceeding having been initiated. A lawsuit is a court proceeding, and filing refers to its initiation, which is a part of the court proceeding. The two words in combination express the "doing" of the SUE event and meet the definition of Justice_Sue. The single verb "sue" can also be used to meet this definition, as can the single noun "lawsuit". However in the third sentence, "contest" is separate from the lawsuit event and does not belong to the event nugget. To contest a lawsuit is an action of the defense team in response to an existing lawsuit. There is currently no Justice Subtype defined in the *Light ERE Guidelines* to fit this contest event.

3.3 REALIS

In our annotation, event nuggets are annotated with three types of REALIS: ACTUAL, GENERIC, and OTHER. REALIS relates to whether or not an event occurred (LTI, 2014).

The REALIS of ACTUAL is used when the event actually happened at a particular place and time, involving specific entities. Both ongoing events and events that have ended are tagged ACTUAL. For example, "He **emailed** her about their plans [Contact_Communicate, ACTUAL]."

The REALIS of GENERIC is used for events that refer to general events involving types or categories of entities. GENERIC is also used for taggable event nuggets which appear in statistics or demographic information. For example, "People **die** [Life_Die, GENERIC]."

The REALIS of OTHER will be used for events that are neither ACTUAL nor GENERIC. If it is determined that an event meets the definition of a type/subtype and it is not an ACTUAL or GENERIC event, it can simply be tagged OTHER. For example, "He plans to **meet** with both political parties [Contact_Meet, OTHER]."

In the case of GENERIC events which also qualify as OTHER (e.g., negated generic) or ACTUAL (e.g., past generic, habitual generic), GENERIC is used, not OTHER or ACTUAL.

4 Event Types/Subtypes

The *TAC KBP Event Guidelines* and the *Light ERE Guidelines* share the same 33 event types/subtypes in particular areas, such as Life, Movement, Business, Conflict, Personnel, Transaction, and Justice, which were originated in the *ACE Guidelines* (LDC, 2005).

The complete set of event types/subtypes is: Life (Be-Born, Marry, Divorce, Injure, Die), Movement (Transport-Person), Business (Start-Org, End-Org, Declare-Bankruptcy, Merge-Org), Conflict (Demonstrate, Attack), Contact (Meet, Communicate), Personnel (Start-Position, End-Position, Nominate, Elect), Transaction (Transfer-Ownership, Transfer-Money), Justice (Arrest-Jail, Release-Parole, Trial-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon).

- John Doe was **born** in Casper, WY. [Life_Be-Born, ACTUAL]
- Roosevelt and his family immediately **departed** for Buffalo. [Movement_Transport-Person, ACTUAL]
- A car bomb **exploded** in central Baghdad. [Conflict_Attack, ACTUAL]

5 Annotation Challenges

One of the main challenges in the development of annotation guidelines is that there is always some disagreement about what should (not) be taggable. In this section, we present some examples of disagreements, which we experienced in the process of developing annotation guidelines, as case studies.

The first case is related to annotating implied events which are contained within nouns referring to persons (e.g., "protestor", "assailant", "killer"). The second case concerns prepositional phrases (e.g., "in prison", "behind bars"), which seem to meet the definitions of event types/subtypes. The third case involves annotating nouns that refer to the consequences or results of events (e.g., "injury", "body", "funeral"), which could be considered as either an entity or an event by individual annotators. The fourth case occurs when only a portion of a word indicates an event (e.g., "antiwar", "post-war", "ex-husband", "ex-wife"). The last case is discontinuous tagging of event nuggets. Although discontinuous tagging is effective for capturing the semantically meaningful unit of event nuggets, the consistency (See Table 5) of discontinues event nuggets is not as good as singe token event nugget.

In the case studies below, the words in ***italic bold*** are controversial or in issue.

Case Study 1: Is a person an event?

- Two other ***assailants*** have committed suicide.
- Here is the KICKER: As reported by local news stations DOZENs of ***protestors*** showed up to protest.
- On the grounds of legality, according to the Geneva Convention, members of regular armed forces – involved in conflicts – are the only persons who may be considered lawful ***combatants*** and authorized to use lethal force.

The words such as "assailants", "protesters", and "combatants" imply the occurrence of events, as we can see by paraphrasing them as "a person who assailed (assaulted) someone," "people who are protesting," and "people who combat." If annotators take the implied occurrences into consideration, those words will be tagged as event nuggets. However, those words actually refer to the "people" themselves. People are not events. Tagging them as events means that we tag implied events. In a similar fashion, some annotators may be tempted to tag "the **dead**" as an event nugget, but others do not because they think that "the dead" refers to dead people. It is critical for annotators to consider the implications of implied events when they tag. If implied events are to be tagged, rules should be explicitly stated to guide annotators as to which implied events should be tagged, and which implied events should not be tagged.

Case Study 2: Is a prepositional phrase taggable?

- A former militant of the French far-left group Action Directe, Georges Cipriani, left prison on parole on Wednesday after 23 years ***behind bars*** for two high-profile murders.

- Prosecutors have said Chen could face life ***in prison*** if convicted on all counts, including embezzlement and bribe-taking."

The phrases "behind bars" and "in prison" indicate that the agent was (or would be) imprisoned and could be tagged as Justice_Arrest-Jail events. They are, however, prepositional phrases that describe a certain state (i.e. the state of physically residing in a particular place). There is some debate whether or not states are taggable as events. Especially in the case of prepositional phrases, it is difficult for annotators to decide whether those phrases should be tagged, since they could be considered to refer to states and sound less eventive.

Case Study 3: Is it an event or the result of an event?

- Why was Trayvon's ***body*** laying 12 hours in the Morgue?
- A cry for the men to be hanged went up almost immediately after the woman died of her ***injuries***, …
- And those already existing time place and manner restrictions were utilized at Matthew Snyder's ***funeral***, with the result that the family never even knew WBC was there.

The words in ***italic bold*** indicate the consequence or result of certain events. For example, the type of "body" referred to in the first example only exists after a Life_Die event has occurred. "Injuries" exist on or in a person's body after (s)he has experienced a Life_Injure or Conflict_Attack event. A "funeral" is a ceremony that occurs after a Life_Die event has happened. Since "body", "injuries" and "funeral" are words that are closely related to taggable event types/subtypes, annotators may be tempted to tag those words as event nuggets. However, it is necessary to differentiate the consequence/result of an event from an event itself.

Case Study 4: Is a portion of a word taggable?

- U.N. Secretary General Kofi Annan said this week that the body has no interest in policing a ***postwar*** Iraq, …
- We were so proud of forming an ***antiwar*** bloc with France and Germany …
- Jurassic Park creator Crichton agrees to pay ***ex-wife*** 31 million dollars

The decision on whether a portion of a word should be tagged also causes disagreement among annotators. Some annotators may think it not appropriate to break a word into chunks, or others may tag a part of a word only if it is hyphenated. This case study raises the issue on how events are defined in relation to word level structure. Semantically, both "war" and "ex" meet the definitions of event types/subtypes. However, it is unclear whether the entire word ("postwar", "antiwar", "ex-wife") should be tagged. Is "antiwar" a Conflict_Attack event, for instance? It is necessary to have a clear rule for this type of tagging.

Case Study 5: Tagging Discontinuous Multiword Event Nuggets

In our corpus with 3,798 event nuggets, there were 209 discontinuous nuggets, a ratio of 5.5%. The discontinuous event nuggets appear in various forms such as verb+noun, verb+particle/adverb, verb+adjective, and verb+prepositional phrase. Among those patterns, the most frequent one is a verb+noun compound (83%), where a noun is the direct object of the verb. This pattern appears in a passive form as well.

- today I **got** a **letter** from the hospital [Contact_Communicate, ACTUAL]
- where was the father when the **shot** was **fired** not more than a 1000 feet away? [Conflict_Attack, ACTUAL]

These discontinuous events are tagged because multiple words in the sentence are important semantic components of their event type/subtype definitions. For example, the word, "get" is used to create various event types such as "get money" (Transaction_Transfer-Money) and "get a job" (Personnel_Start-Position). Thus, tagging a verb and a noun together as one event seems important to differentiate a particular event type from the others. In the second example, both "shot" and "fired" are taggable as events and it is hard to ignore either of them as not taggable due to the close relationship between the "doing" of an event and event itself. A verb+noun compound appears very often in the following event types/subtypes: Transaction_Transfer-Money (23%), Contact_Communicate (18%), and Conflict_Attack (10%).

Part of speech patterns for discontinuous tagging include verb+particle/adverb, which is 14% of the entire discontinuous tagging. This form appears most often in Movement_Transport-Person (68%).

- …**took** us **in** for a interview…[Movement_Transport-Person, ACTUAL]
- ... i **put** the thread **up** because i really did want some opinions…[Contact_Communicate, ACTUAL]

Some annotators may only tag main verbs because they think adverbs and particles are modifying the verbs, but others may tag verb+adverb/particles together because they feel that the adverb/particles signify a different meaning from just the verbs alone. As shown Table 5, it is not as easy to consistently annotate multi-word event nuggets as it is to consistently annotate single-word event nuggets. However, the percentage of multi-word event nuggets is so low that it may not significantly affect overall event nugget detection performance.

We continue to work on reaching agreement on the optimal method of handling of these four types of controversial event nuggets in order to better represent the deeper semantics of texts. The very low frequency of discontinuous event nuggets does not mean that they should be ignored to achieve higher inter-annotator agreement. Clear rules for these cases should be laid out for future tasks on event nugget detection.

6 Brat Rapid Annotation Tool (brat)

Our annotation was conducted using Brat Rapid Annotation Tool (brat). This tool allows for customization of tags, such as event types/subtypes, realis types, types of entities/arguments, types of event links, and provides a means to add notes for questionable mentions. In addition, brat supports discontinuous tagging and side-by-side comparison of two files.

The actual procedure of annotation and the review of applied tags are relatively simple with this user-friendly application. Clicking on a word to be tagged opens a window where annotators can select tags, such as event types/subtypes and realis. After a word has been tagged, when the cursor is moved over the tag, a small box appears, displaying the assigned event type and realis for review. Screenshots of brat are shown in the Appendix.

7 Data Selection and Preparation

We produced training and evaluation (eval) data to support the Event Nugget evaluation as a pilot TAC KBP evaluation. The data includes both formal newswire text (NW) and informal discussion forums (DF), drawn from a pool of data also labeled for the DARPA DEFT Program's Light Entities, Relations and Events (Light ERE) task (Song et al., 2015), and/or the NIST TAC KBP Evaluation Event Argument Task (Ellis et al., 2014), with the goal of ultimately being able to take advantage of multiple styles of event annotation on the same data. Documents for the current task were carefully selected from this pool to optimize coverage of as many of the event types and subtypes as possible, with a goal of at least five instances of each type-subtype combination. The training data consists of 151 documents, while the eval data contains 200 documents. Table 1 shows the genre distribution as well as token counts for each partition.

Partition	Training		Eval	
Genre	NW	DF	NW	DF
Documents	77	74	101	99
Tokens	44,962	70,427	50,997	169,740

Table 1. Event Nugget Data Profile

While the Light ERE and KBP Event Argument tasks rely on character offsets for annotation and scoring, the Event Nugget Tuple Scorer[2] (Liu, Mitamura & Hovy, 2015) requires tokenized data. Therefore, prior to annotation, all selected documents were automatically tokenized in the Penn English Treebank style. No manual correction was performed on the tokenization due to time constraints.

8 Corpus and Consistency Analysis

8.1 Corpus

Experience with event annotation for Light ERE and ACE (Doddington et al., 2004) and related tasks suggests that a major challenge for annotation consistency is poor recall – human annotators are not highly consistent in recognizing that a mention has occurred. To reduce the impact of this known issue for the Event Nugget task, two anno-

[2] Event Nugget Tuple refers to the tuple made up of the nugget, event type/subtype, and realis.

tators independently labeled each document (two first pass annotation passes, referred to as FP1 and FP2 below); a senior annotator then adjudicated discrepancies to create a gold standard. The team consisted of four first pass annotators, two of whom were also adjudicators. The effort was made to ensure that annotators did not adjudicate their own first pass files, but due to time constraints and the pilot nature of the task, in some cases there was overlap.

The gold standard training data has 3,798 event nuggets annotated in total, while the eval data has 6,921 event nuggets. Table 2 shows the distribution of event nuggets by genre and realis type for each partition.

Realis Attribute	Training		Eval	
	NW	DF	NW	DF
Generic	202	383	245	981
Other	346	406	448	1271
Actual	1313	1132	1752	2224
Total	3798[3]		6921	

Table 2. Realis Annotation of Event Nuggets

Figure 1 (in Appendix) shows the distribution of each type-subtype combination in the training and eval data. Conflict_Attack has the highest representation in both training (579) and eval (791). Justice_Extradite has the lowest count in training data (3), while Life_Be-Born is least frequent in the eval data (19). Despite our efforts to manually select documents to maximize coverage for all type-subtype combinations, the corpus does not include any occurrences of Business_End-Org or Personnel_End-Position.

8.2 Consistency Analysis

We examined annotation consistency and quality by comparing different passes of the eval set annotation using the Event Nugget Tuple Scorer (Liu, Mitamura, & Hovy, 2015) developed for the event nugget evaluation task. This scorer treats one file as “gold” and the other as “system”, and matches each nugget in the gold file to one or more nuggets in the system file. This mapping is based on the overlap of the nugget spans. By nugget span, we mean the exact list of tokens, continuous or discontinuous, that make up an event nugget. However, each system nugget can only be mapped to one gold nugget. For each gold nugget, the scorer computes type and realis accuracy scores based on the values for the gold nugget and all the system nuggets that are mapped to it.

The scorer produces three scores for each file. The first is an F-measure for the nugget spans, based on the mapping from gold to system nuggets, as well as ‘‘false alarms” in the system file that are not mapped to any nuggets in the gold file. The type and realis scores for each gold mention are also cumulatively summed up, producing a type and realis score for the file. The type and realis scores are therefore tied to the F-measure score of the nugget spans. We used this scorer rather than the ACE (NIST, 2005) scorer since this scorer was designed for the event nugget evaluation task, and so seemed the most appropriate to use for evaluation of annotation consistency and quality of this corpus.

We examined annotation consistency by comparing the two independent first passes of annotation (FP1 and FP2), with the results shown in the column FP1 vs. FP2 in Table 3. We also evaluated improvement in annotation quality in the workflow by comparing the adjudicated (ADJ) and first (FP1 and FP2) passes, shown in the columns ADJ vs. FP1 and FP2 in Table 3. The noticeable improvement in score shows the advantage of including adjudication as part of the annotation process. (For IAA purposes, there is obviously no gold or system, but in order to use the scorer we arbitrarily treated one file as the “gold”.)

	FP1 vs. FP2	ADJ vs. FP1	ADJ vs. FP2
Span	69.0	78.2	89.3
Type	68.2	71.7	84.3
Realis	60.0	63.2	85.7

Table 3. Scores for Event Nugget Eval Set Annotation

To gain some further insight into these numbers we expanded the analysis in two directions. First, we compared the FP1 vs. FP2 event nugget consistency with the FP1 vs. FP2 annotation consistency on the ACE 2005 training data (Walker et al., 2006). There is also a scorer that was developed for ACE (NIST, 2005), but we used the Event Nugget Tuple evaluation scorer so that we could score both sets of data for this comparison as in the

[3] 16 event nuggets in the training set did not receive a realis attribute, due to annotation error.

event nugget evaluation. This necessitated converting the ACE files into the format for event nuggets used for the current scorer. We used the ‘‘anchor’’ string of the ACE event mention as the nugget span, the ‘‘type’’ and ‘‘subtype’’ of the ACE event mention as the nugget type, and the ‘‘modality’’ of the ACE event mention as the nugget realis value. The results are shown in Table 4. The ACE FP1 vs. FP2 scores in Table 4 are somewhat lower than the FP1 vs. FP2 scores for the event nugget annotated data. However, while we have converted the format and used the same scorer, the annotation task is not identical, so this can only be taken as a rough comparison. There is greater difference between the ADJ vs. FP1, FP2 scores for the event nugget data than the ACE data. The event nugget task had a smaller annotation team than for ACE, and it is likely that more of the adjudication annotators for event nugget annotation also did the FP2 pass than was the case for ACE.

	FP1 v. FP2	ADJ v.FP1	ADJ v. FP2
Span	64.8	79.3	81.8
Type	62.2	70.4	75.6
Realis	56.1	68.0	73.0

Table 4. Scores for ACE 2005 Training Annotation

Second, we wished to determine also if there was a difference in the annotation consistency and quality of event nugget spans depending on whether the span consists of only one token as compared to those that are multiple tokens, either continuous or discontinuous. We decomposed the span F-measure in Tables 3 and 4 based on these criteria. We did this by modifying the event nugget scoring program to optionally ignore nuggets depending on their span. For example, when we wished to compare annotations for which the span is a single token, we simply ignored all nuggets with spans of more than one token. Likewise, when comparing nuggets for which the span consists of discontinuous multiple tokens, all nuggets for which the span was either a single token or multiple continuous tokens were ignored.

We ran this modified scorer in different modes to use (1) all nuggets (as before), (2) only nuggets that consist of a single token, ignoring all others, (3) only nuggets that consist of multiple continuous tokens, (4) only nuggets that consist of multiple discontinuous tokens, and (5) only nuggets that consist of multiple tokens, whether continuous or not. Mode (1) is the same as the score reported for the spans in Tables 3 and 4, and modes (2)-(5) in effect break this down into subcomponents. The results are shown in Table 5. ACE annotation did not allow discontinuous multiple token mentions, and so there are no results listed for ACE for (4) and (5).

The results for the consistency agreement between FP1 and FP2 show a similar fall in score for both the event nugget data and the ACE 2005 training data, when considering only multiple continuous tokens. The score climbs back up a little for the event nugget FP1 vs. FP2 score when considering (5) either continuous or discontinuous multiple tokens, as compared with either (3) only multiple continuous or (4) only multiple discontinuous. The reason for this is that there are cases where one file has an event nugget with a continuous multiple token span such as “got jail time” while the other has the corresponding event nugget with a multiple discontinuous span such as ‘‘got time’’. In (3) or (4), only one or the other would be included in the comparison, whereas in (5) and (1) both would be included, allowing for partial match instead of a miss. Similarly, there are cases where one file has a single token span for a nugget while the other file has a multiple token span for the corresponding nugget, and so it is only in (1) that both would be included, allowing for a partial match instead of a miss.

These more fine-grained nugget span scores for FP1 vs. FP2 show that single-token nuggets are annotated more consistently than multi-token nuggets. Considering just the multi-token nuggets, there is little difference in consistency of annotation between continuous and discontinuous spans. The ADJ vs. FP1 / ADJ vs. FP2 results show that including adjudication annotation lessens any difference in annotation quality for nuggets depending on whether the span is single or multi-token.

In future work on this consistency analysis, we will also go in the other direction, and convert the event nugget data into the ACE format so that it can be evaluated using the ACE scorer (NIST, 2005), ensuring that the comparison of inter-annotator consistency is not overly affected by details of particular scoring algorithms.

	Event Nugget				ACE 2005 Training			
	FP1 vs. FP2		ADJ vs. FP1 / ADJ vs FP2*		FP1 vs. FP2		ADJ vs. FP1 / ADJ vs. FP2	
	Span F-meas	Ratio**	Span F-meas	Ratio	Span F-meas	Ratio	Span F-meas	Ratio
(1) All mentions	69.0	100%	78.2/89.3	100%	64.9	100%	79.3/81.8	100%
(2) Single-token	67.7	90.0%	77.0/88.9	87.7%	65.0	94.6%	79.2/81.6	95.2%
(3) Multiple cont.	45.3	6.1%	57.7/84.4	6.8%	44.2	5.4%	70.8/70.6	4.8%
(4) Multiple discont.	43.0	4.0%	57.5/84.1	5.5%	NA	NA	NA	NA
(5) Multiple all	46.0	10.1%	59.0/85.4	12.3%	NA	NA	NA	NA

Table 5: Decomposing the Span Scores for Nugget and Trigger Span

* The two figures represent ADJ compared to FP1 (before the slash) and ADJ compared to FP2 (after the slash).

** Event nugget type per all event nuggets.

9 Conclusion

This paper first describes the processes of event nugget annotation using a brat tool and issues which arose in the process of developing *TAC KBP Event Guidelines*. We present complex cases that cause annotators' disagreement on tagging. Questions are raised about implied events, states vs. events, results of events, tagging portions of words, and discontinuous tagging. Second, the paper explains the creation of a tagged event nugget corpus and provides annotation statistics and consistency analysis comparing the first pass annotations, and also a comparison of adjudicated files with first pass files using the Event Nugget Tuple Scorer. The analysis shows that single-word nuggets are tagged more consistently than multi-word nuggets and that adjudication is very important for improving the quality of annotation.

Reconciliation of annotation disagreement is crucial in terms of not only the development of annotation guidelines but also the quality of annotation. This is closely associated with how an event nugget is defined and clarification of tagging rules. Resolving the issues surrounding event type/subtype definitions will be very helpful not only for future studies on event nugget detection but also studies on event coreference, subevent relations, and event arguments.

Acknowledgments

This material is based on research sponsored by Air Force Research Laboratory and Defense Advanced Research Projects Agency under agreement numbers FA8750-13-2-0045 and FA8750-12-2-0342. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and Defense Advanced Research Projects Agency or the U.S. Government.

We are thankful to Jun Araki, Zhengzhong “Hector” Liu, and Volkan Cirik for their assistance in the annotation tool and data calculation.

References

George Doddington, Alexis Mitchell, Mark Przbocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, May 24-30.

Joe Ellis, Jeremy Getman, and Stephanie M. Strassel. 2014. Overview of Linguistic Resources for the TAC KBP 2014 Evaluations: Planning, Execution, and Results. In *Proceedings of TAC KBP 2014 Workshop*, National Institute of Standards and Technology, Gaithersburg, Maryland, USA, November 17-18, 2014.

Language Technologies Institute. 2014. *TAC KBP Event Detection Annotation Guidelines, Version 1.7*, Language Technologies Institute, CMU, September 12, 2014.

Linguistic Data Consortium. 2014. *DEFT ERE Annotation Guidelines: Events Version 1.3*, March 13, 2014.

Zhengzhong Liu, Teruko Mitamura, Eduard Hovy. 2015. "Evaluation Algorithms for Event Nugget Detection: A Pilot Study". To appear in the Proceedings of the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation. NAACL-HLT 2015.Linguistic Data Consortium. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events, Version 5.4.1 2005.05.09*.

National Institute of Standards and Technology. 2005. *The ACE 2005 Evaluation Plan*. http://www.itl.nist.gov/iad/mig/tests/ace/2005/doc/ace05-evalplan.v3.pdf

Zhiyi Song, Ann Bies, Tom Riese, Justin Mott, Jonathan Wright, Seth Kulick, Neville Ryant, Stephanie Strassel, Xiaoyi Ma. Submitted. From Light to Rich ERE: Annotation of Entities, Relations, and Events.

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 the Demonstrations Session at EACL 2012*.

Christopher Walker, Stephanie Strassel, Julie Medero, Kazuaki Maeda. 2006. *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium Catalog No.: LDC2006T06.

Appendix

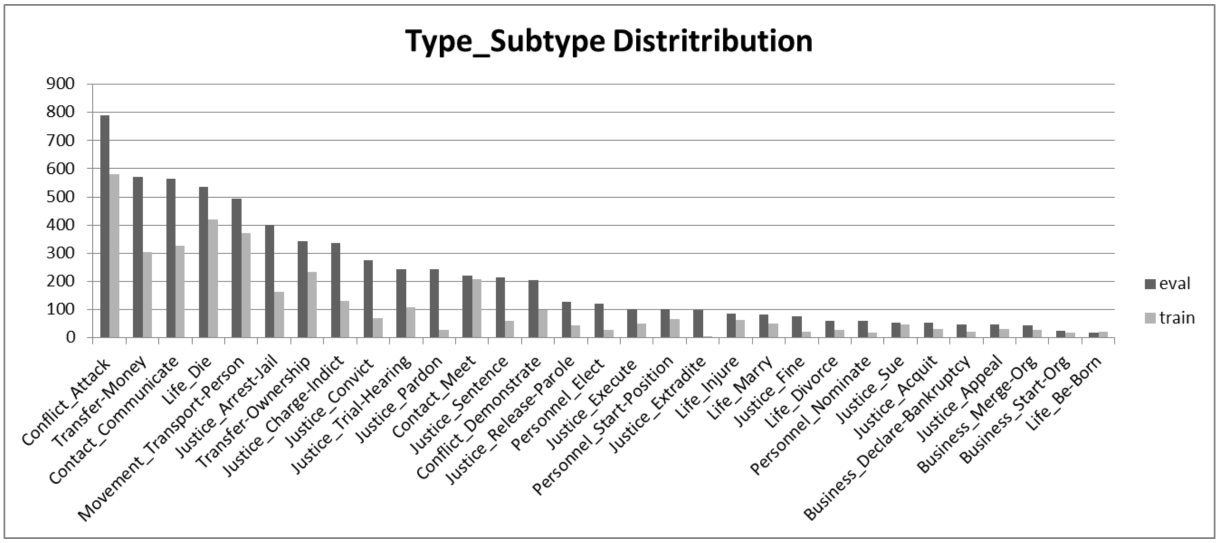
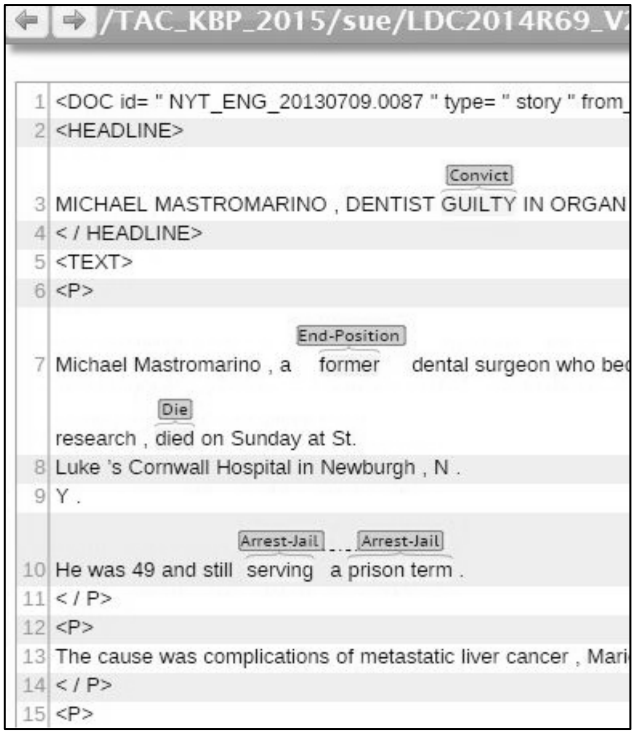


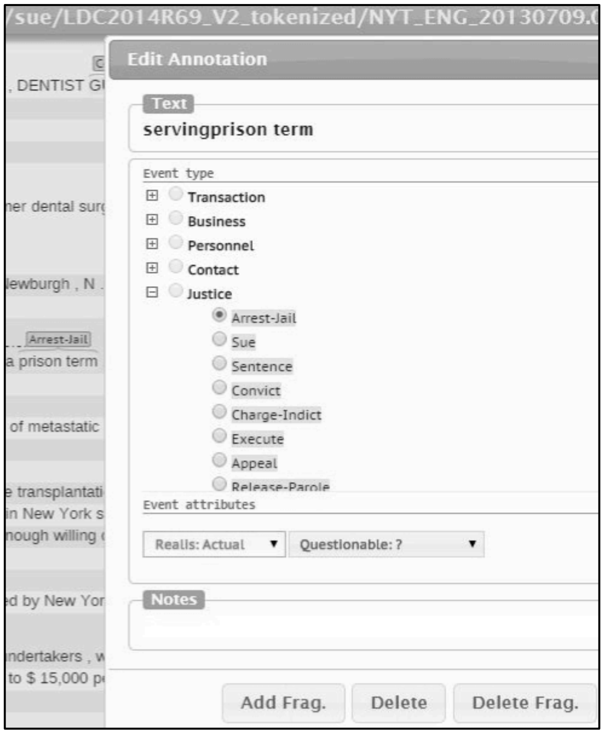
Figure 1. Type and Subtype Distribution in Event Nugget Annotation



/TAC_KBP_2015/sue/LDC2014R69_V2

<DOC id= " NYT_ENG_20130709.0087 " type= " story " from_
<HEADLINE>
Convict
MICHAEL MASTROMARINO , DENTIST GUILTY IN ORGAN
< / HEADLINE>
<TEXT>
<P>
End-Position
Michael Mastromarino , a former dental surgeon who bec
Die
research , died on Sunday at St.
Luke 's Cornwall Hospital in Newburgh , N .
Y .
Arrest-Jail Arrest-Jail
He was 49 and still serving a prison term .
< / P>
<P>
The cause was complications of metastatic liver cancer , Mari
< / P>
<P>

Screenshot 1. Brat tool main annotation screen



/sue/LDC2014R69_V2_tokenized/NYT_ENG_20130709.0

Edit Annotation

Text
servingprison term

Event type
○ Transaction
○ Business
○ Personnel
○ Contact
○ Justice
● Arrest-Jail
○ Sue
○ Sentence
○ Convict
○ Charge-Indict
○ Execute
○ Appeal
○ Release-Parole

Event attributes
Realis: Actual
Questionable: ?

Notes

Screenshot 2. Brat tool pop-up window