Event Nugget and Event Coreference Annotation

Zhiyi Song[a], Ann Bies[a], Stephanie Strassel[a], Joe Ellis[a], Teruko Mitamura[b], Hoa Dang[c], Yukari Yamakawa[b], Sue Holm[b]

[a]Linguistic Data Consortium, University of Pennsylvania,
3600 Market Street, suite 810, Philadelphia, PA 19104, USA
{zhiyi, bies, strassel, joellis}@ldc.upenn.edu
[b]Language Technologies Institute, Carnegie Mellon University,
6711 Gates-Hillman Center, 5000 Forbes Ave., Pittsburgh, PA 15213, USA
teruko+@cs.cmu.edu, yukariy@andrew.cmu.edu, sh4s@andrew.cmu.edu
[c]National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA
hoa.dang@nist.gov

Abstract

In this paper, we describe the event nugget annotation created in support of the pilot Event Nugget Detection evaluation in 2014 and in support of the Event Nugget Detection and Coreference open evaluation in 2015, which was one of the Knowledge Base Population tracks within the NIST Text Analysis Conference. We present the data volume annotated for both training and evaluation data for the 2015 evaluation as well as changes to annotation in 2015 as compared to that of 2014. We also analyze the annotation for the 2015 evaluation as an example to show the annotation challenges and consistency, and identify the event types and subtypes that are most difficult for human annotators. Finally, we discuss annotation issues that we need to take into consideration in the future.

1 Introduction

The Text Analysis Conference (TAC) is a series of workshops organized by the National Institute of Standards and Technology (NIST), aiming to encourage research in natural language processing (NLP). The Knowledge Base Population (KBP) tracks of TAC encourage the development of systems that can match entities mentioned in natural texts with those appearing in a knowledge base and extract information about entities from a document collection and add it to a new or existing knowledge base. Starting in 2014, TAC KBP added a track for event evaluation. The goal of the TAC KBP Event track is to extract information about events such that the information would be suitable as input to a knowledge base.

Event Nugget (EN) evaluation, as one of the evaluation tasks in the TAC KBP event track, aims to evaluate system performance on EN detection and EN coreference (Mitamura & Hovy, 2015). An event nugget, as defined by the task, includes a text extent that instantiates an event, a classification of event type and subtype, and an indication of the realis of the event. This is different from nuggets (Voorhees, 2003; Babko-Malaya, et al, 2012) and the Summary Content Units (SCU) described in (Nenkova and Passnoneau, 2004; Nenkova et al, 2007) where nuggets and SCU are units of meaning (usually in the form of a sentential clause) which not only includes the main verb/word that instantiates an event, but also all arguments.

In this paper, we describe the event nugget annotation to support the pilot EN Detection evaluation

Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation, pages 37–45,
San Diego, California, June 17, 2016.

in 2014 and in support of the EN Detection and Coreference TAC KBP open evaluation in 2015. We discuss changes to the annotation in 2015 as compared to that of 2014 and present the data volume annotated for both training and evaluation data for the 2015 evaluation. We analyze the annotation for the 2015 evaluation data to show the annotation and consistency challenges, identify the event types and subtypes that are most difficult for human annotators, and finally discuss annotation issues that we need to take into consideration in the future.

2 Event Nugget Evaluation in TAC KBP

The EN evaluation was introduced in 2014 and run as a pilot evaluation, which sought to serve two purposes. One was to measure event detection by performers' systems in the Deep Exploration and Filtering of Text (DEFT) program of the Defense Advanced Research Projects Agency (DARPA, 2012). The program aims to address remaining capability gaps in state-of-the-art natural language processing technologies related to inference, causal relationships and anomaly detection. The other purpose of the pilot, however, was to test run the evaluation framework before opening it up to the full TAC KBP community. The Pilot EN evaluation had one evaluation task:

- EN Detection: Participating systems must identify all relevant Event Mention instances in English texts, categorize each mention's Event types/subtype and identify its realis value (ACTUAL, GENERIC, or OTHER).

As the pilot was considered a success, EN was added to the roster of full-fledged evaluation tracks for TAC KBP 2015, with some modifications to incorporate lessons learned from the pilot and to better align with the other TAC KBP 2015 event-related evaluations – Event Argument Linking (EAL) (Freedman, 2014 & 2015) – and also the Entities, Relations and Events (ERE) data provided to KBP participants for training purposes. The EN evaluation in TAC KBP 2015 included two new tasks in addition to EN Detection:

- EN Detection and Coreference: Participating systems must identify not only the event nugget, but also full event coreference links. Full Event Coreference is identified when two or more event nuggets refer to the same event.
- EN Coreference: Participating systems must identify full event coreference links, given the annotated event nuggets in the text.

ERE was developed as an annotation task that would be supportive of multiple research directions and evaluations in the DEFT program, and that would provide a useful foundation for more specialized annotation tasks like inference and anomaly. The resulting ERE annotation task has evolved over the course of the program, from a fairly lightweight treatment of entities, relations and events in text (Light ERE), to a richer representation of phenomena of interest to the program (Rich ERE) (Song, et al., 2015). In ERE Event annotation, each event mention has annotation of event type and subtype, its realis attribute, any of its arguments or participants that are present, and a required "trigger" string in the text; furthermore, event mentions within the same document are coreferenced into event hoppers (Song, et al., 2015). EN annotation includes all of these annotations, except the annotation of event arguments.

The EN task in 2014 adapted the event annotation guidelines from the Light ERE annotation task (Aguilar, et al., 2014) by incorporating modifications by the evaluation coordinators that focused on the text extents establishing valid references to events, clarifications on transaction event types, and the additional annotation of event realis attributes, which indicated whether each event mention was asserted (Actual), generic or habitual (Generic), or some other category, such as future, hypothetical, negated, or uncertain (Other) (Mitamura, et al., 2015) .

In 2015, EN annotation followed the Rich ERE Event annotation guidelines (except for the annotation of event arguments). As compared to EN annotation in 2014, Rich ERE Event annotation and 2015 EN annotation include increased taggability in several areas: slightly expanded event ontology, additional attributes for contact and transaction events, and double tagging of event mentions for multiple types/subtypes and for certain types of coordination, in addition to event coreference.General Instructions

3 Event Nugget Annotation

In this section, we describe the EN annotation as well as the major differences between the 2014 and 2015 annotation tasks.

3.1 Event Trigger

An event trigger span is the textual extent within a sentence that indicates a reference to a valid event of a limited set of event types and subtypes. In the 2014 EN task, the trigger span was defined as a semantically meaningful unit which could be either a single word (main verb, noun, adjective, adverb) or a continuous or discontinuous multi-word phrase (Mitamura et al., 2015). For the 2015 EN evaluation, trigger span was redefined as the smallest, contiguous extent of text (usually a word or phrase) that most saliently expresses the occurrence of an event. This change brings consistency to the EN annotation of event trigger with approaches taken in Automatic Content Extraction (ACE) (LDC, 2005) as well as Light and Rich ERE (Song et al, 2015). Additionally, we see improved annotation consistency in terms of event trigger extent, as shown in Figure 1 of section 5.1. Unlike in the 2014 EN annotation, annotators for the 2015 data were allowed to 'double tag' event triggers in order to indicate that a given text extent referred to more than one event type/subtype, which was usually used to indicate the presence of an obligatorily inferred event. Double tagging was also used for certain types of coordination. For example, given the following text:

> *Cipriani was sentenced to life in prison for the **murder** of Renault chief George Besse in 1986 and the head of government arms sales Rene Audran a year earlier.*

In 2015 EN annotation, the word "murder" would be the trigger for two Life.Die events, one with the victim "George Besse" and the other with "Rene Audran" as well as two Conflict.Attack events, one occurring in 1986 and the other in 1985. In 2014 EN annotation, the word "murder" would be the trigger for only one Conflict.Attack event.

3.2 Event Taxonomy

EN annotation and evaluation focus on a limited inventory of event types and subtypes, as defined in ERE, based on Automatic Content Extraction (Doddington et al., 2004; Walker et al., 2006; Aguilar et al., 2014; Song et al., 2015). The 2014 EN evaluation covered the inventory of event types and subtypes from Light ERE, including 8 event types and 33 subtypes.

The 2015 evaluation added a new event type (Manufacture) and four new subtypes – Movement.TransportArtifact, Contact.Broadcast, Contact.Contact, Transaction.Transaction – which aligned the EN event ontology with that of Rich ERE in order to take advantage of the existing Rich ERE annotated data as training data. The EN annotation task also adopted a new approach for applying the Contact event subtype categorizations, which had been developed for Rich ERE data creation efforts. Instead of having annotators categorize the subtypes directly, Contact event mentions were labeled with attributes to describe formality (Formal, Informal, Can't Tell), scheduling (Planned, Spontaneous, Can't Tell), medium (In-person, Not-in-person, Can't Tell), and audience (Two-way, One-way, Can't Tell). Contact event subtypes were automatically generated based on the annotated attributes. This change added increased granularity to the Contact events and captured more subtypes, which had been requested by data users, and it also allowed the annotation to provide information at two levels, the attribute level and the traditional event subtype level, which may support more robust system development.

3.3 Event Coreference

The final change to the 2015 EN evaluation as compared to the 2014 pilot was the added requirement of event coreference. Again following the Rich ERE task, EN annotation in 2015 adopted the notion of 'event hoppers', a more inclusive, less strict notion of event coreference as compared to previous approaches as in ACE (LDC, 2005) and Light ERE. The notion of event hopper was introduced to address the pervasive challenges of event coreference, with respect to event mention and event argument granularity (Song, et al., 2015). Following this approach, event mentions were added to an event hopper when they were intuitively coreferential to an annotator, even if they did not meet a strict event identity requirement. Event nuggets could be placed into the same event hoppers even if they differed in temporal or trigger granularity, their arguments were non-coreferential or conflicting, or even if

their realis mood differed, as long as they referred to the same event with the same type and subtype. For example, in the following two sentences:

- *The White House didn't confirm Obama's* ***trip*** ***{Movement.TransportPerson, Other}*** *to Paris for the Climate Summit last year.*
- *Obama* ***went*** ***{Movement.TransportPerson, Actual}*** *to Paris for the Climate Summit.*

The first event nugget's realis label is Other while the second event nugget is Actual. From the context, we know that both nuggets are talking about the same trip to Paris, so even though they have different realis labels, they still belong to the same event hopper.

4 Training and Evaluation Data

In this section, we present the training and evaluation data annotated to support the EN evaluation in 2015.

4.1 Training Data

Due to the changes in Rich ERE event annotation (and hence 2015 EN annotation) as compared to EN annotation in 2014, the 2014 evaluation data set was re-annotated so that the annotation matched the Rich ERE standard. Additionally Rich ERE annotation is created as a core resource for the DEFT program and TAC KBP evaluation, aiming to provide a valuable resource for multiple evaluations. Rich ERE annotation includes exhaustive annotation of Entities, Relations, Events and Event Hoppers, and 2015 EN annotation shares the Rich ERE annotation guidelines for Events, with the exception that Events and Event hoppers in Rich ERE also include the annotation of event arguments. Table 1 lists the total training data volume available for the 2015 EN evaluation.

Annotation	Genre	Files	Words	EN	Hoppers
EN and Coref	NW	81	27,897	2,219	1,461
EN and Coref	DF	77	97,124	4,319	1,874
Rich ERE	DF	240	156,041	4,192	3,044
Rich ERE	NW	48	23,999	1,571	1,099
Total		446	305,061	12,301	7,478

Table 1: Training Data Volume for 2015 Event Nugget.

4.2 Evaluation Data

Source data for the 2015 EN evaluation was a subset of the documents selected for EAL evaluation, which had been manually selected to ensure coverage of all event types and subtypes for that evaluation. Tokenization of the source documents was also provided. Unlike the 2014 data, in which annotation was performed on pre-tokenized text, in 2015 tokenization was performed as a post-annotation procedure, using tool kits provided by evaluation coordinators.

In order to reduce the impact of low recall on annotation consistency, which had proven problematic in the pilot and in previous event annotation efforts (Mitamura et al., 2015), gold standard EN data was produced by first having two annotators perform EN annotation (which included the creation of event hoppers) independently for each document (referred to as first pass 1 or FP1, and first pass 2 or FP2, below), which was followed by an adjudication pass conducted by a senior annotator to resolve disagreements and add annotation that was otherwise missed in one of the first passes. The EN annotation team consisted of nine annotators, six of whom were also adjudicators, and care was taken to ensure that annotators did not adjudicate their own files. Following adjudication of all documents, a corpus-wide quality control pass was also performed. In this pass, annotators manually scanned a list of all event triggers to review event type and subtype values and all event hoppers to make sure that event mentions in the same hopper have same type and subtype value. All identified outliers were then manually reviewed in context, and corrected if needed.

Annotation	Genre	Files	Words	EN	Hoppers
EN and Coref	NW	98	49,319	3,788	2,440
EN and Coref	DF	104	39,333	2,650	1,685
Total		202	88,652	6,438	4,125

Table 2: Evaluation Data Volume for 2015 Event Nugget.

The evaluation data set consists of 202 documents with a total word count of 88,652. The gold standard annotation has a total of 6,438 event nuggets and 4,125 event hoppers in total. Table 2 shows the profile of the evaluation dataset.

Appendix 1 shows the distribution of each type-subtype in the evaluation data. Conflict.Attack has the highest representation (591 event nuggets) while

Business.EndOrg has the lowest count (6). With the two newly added contact subtypes (Contact and Broadcast), there are altogether 1,491 contact event nuggets (23%), with 1,101 Contact.Contact and Contact.Broadcast combined (17%). Each event nugget is labeled with one of the three realis attributes: actual, generic and other. Table 3 shows the distribution of event nugget realis annotation by genres.

realis	NW	DF
actual	2,508	1,595
generic	603	539
other	677	516
total	3,788	2,650

Table 3: Event Nugget counts by realis attributes in NW and DF genres.

5 Inter-annotator Agreement and Annotation Challenges

Subsequent analysis of inter-annotator agreement in the EN 2015 evaluation data indicates that several challenges remain to be addressed.

5.1 Inter-annotator Agreement

Annotation consistency is generally in line with what we expect due to the complex nature of event recognition. The changes in the approach to the annotation task that were described above appear to have made some improvements, as shown in Figure 1, which compares the overall inter-annotator agreement (IAA) on first pass EN annotation in 2014 and 2015. Compared with 2014, IAA F1 score in 2015 improved by 5% in event trigger detection ("plain" in Figure 3) and realis attribute labelling. There is only 1% of improvement on event type and subtype classification. Regarding event coreference, which was new in 2015, the IAA F1 score was 67.63%. Below are a few examples indicating disagreement in event trigger, event typing and realis attributes:

Trigger extent mismatch:

- *She met the insurance investment magnate Shelby Cullom Davis on a **train** (FP1: Movement.TransportPerson) **to** (FP2: Movement.TransportPerson) Geneva in 1930.*

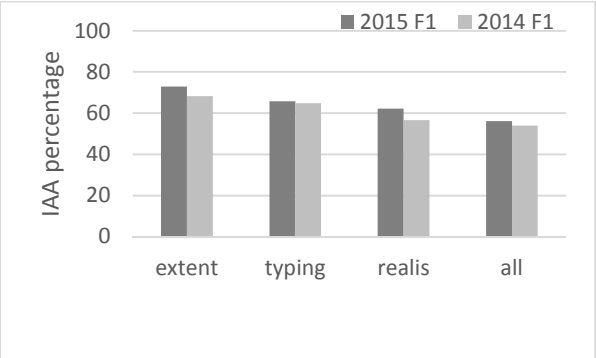


Figure 1: Inter-annotator agreement on first pass EN annotation in 2014 and 2015.

Mention type/subtype mismatch:

- *Dr. Yusuf Sonmez whom he called a notorious international organ **trafficker**. (FP1: Movement.TransportArtifact; FP2: Transaction.TransferOwnership)*

Realis Attribute mismatch:

- *The wealthy, ailing patients who were to **receive** (FP1: Actual; FP2: Other) the organs flew to Pristina.*

Some types scored better than others. As expected, Contact and Transaction event types have the lowest consistency. Figure 2 displays the IAA F1 scores on first pass EN annotation by event types.

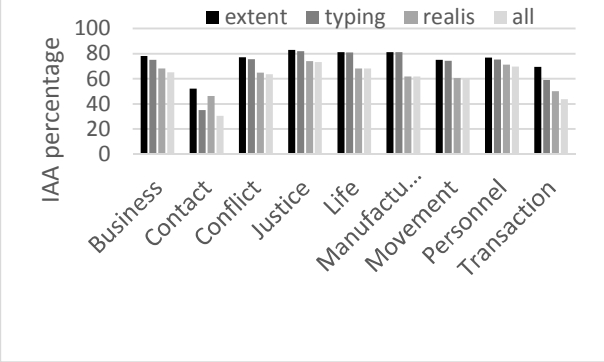


Figure 2: Inter-annotator agreement on first pass EN annotation 2015 by event types

5.2 Taggability

The annotation consistency can be attributed not only to disagreement in terms of event trigger, classification of event type and subtypes, realis attributes, but also misses and false alarms. Determining whether or not an event is taggable (i.e., recall) has always been a difficult issue in event annotation tasks, and some event types and subtypes still ap-

pear to be more heavily affected than others. To better understand the issue, we calculated the level of disagreement for event taggability by dividing the difference of event type-subtype counts in FP1 and FP2 by the event type-subtype count produced in the adjudication pass. Seventeen type-subtypes vary between FP1 and FP2 at a percentage below 10%, including Life.Divorce and Conflict.Attack, with some type-subtypes showing a high level of disagreement. As indicated in Figure 3, there are 12 type-subtypes with 20% or higher disagreement between FP1 and FP2, four of which are over 50%.

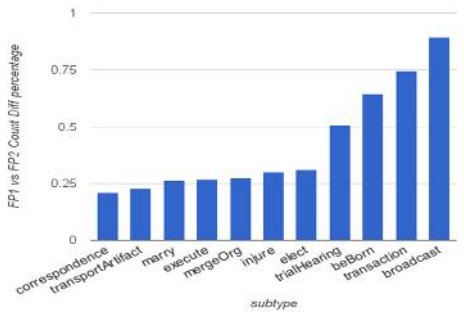


Figure 3: Difference of Event Nugget Occurrence between FP1 and FP2 over Adjudicated Occurrence.

Below are a few examples indicating disagreement in terms of taggability:

Miss:

- where she ***created*** (Business.StartOrg) the Davis Museum and Cultural Center.

False Alarm:

- She also ***owned*** (Transaction.TransferOwnership) a home in Northeast Harbor, Maine

5.3 Contact Event Type

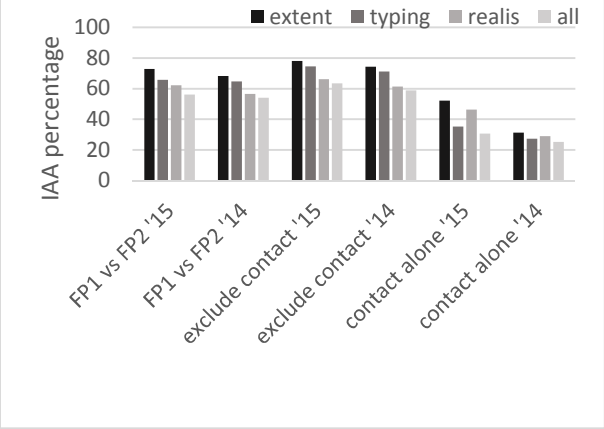


Figure 4: First pass inter-annotator agreement, including and excluding contact events.

One main change that we made in 2015 was the annotation of contact events, aiming to support a wider range of potential subtypes and also to improve annotation consistency. The subtypes were automatically determined based on the annotation of attributes in contact events, rather than by having annotators make a direct decision about contact subtypes.

There are 1479 event nuggets annotated with the contact type, which is 23% of all event nuggets in the evaluation data set. Contact.Broadcast and Contact.Contact are two of the most common subtypes. The consistency of the Contact event type annotation still poses challenges. As indicated in Figure 4, overall annotator consistency in both 2014 and 2015 improve when excluding the contact type from the comparison.

The consistency of contact events alone is much lower as compared with the other event types. This could be attributed to the taggability question of certain verbs. We have specified in the annotation guidelines that speech verbs such as "said" and "told" are taggable event triggers. We also specified that when there were multiple speech verbs instantiating the same contact event, only the first one is taggable. However, annotators vary in implementing this rule, as shown in the following example:

- *A spokesman for the Investigative Committee, a branch of the prosecutor's office,* ***said*** *(FP1: Contact.Broadcast) in an* ***interview*** *(FP2: Contact.Broadcast) in Izvestia.*

One positive thing we can see is that the changes in 2015 did improve the annotation consistency for contact events in 2015, but there is still room for further improvement, especially in the consistency of subtype classification of the Contact event type.

5.4 Double Tagging

Double tagging is quite common in the EN evaluation data set, due to the high frequency of Conflict.Attack/Life.Die, Transaction.TransferMoney and Transaction.TransferOwnership events. Altogether, 575 pairs out of 6440 event nuggets were double tagged (18%). Most of the double tagging cases are the result of the same trigger instantiating different event type-subtypes. Only 12 pairs of double tagging cases are the result of event argument conjunction. Figure 5 shows the distribution of event subtype counts involved in double tagging.

Double tagging was added to address the issue of inconsistency in event type categorization for triggers that may instantiate two or more event types and subtypes. Further analysis is needed to show whether allowing double tagging improves annotation consistency.

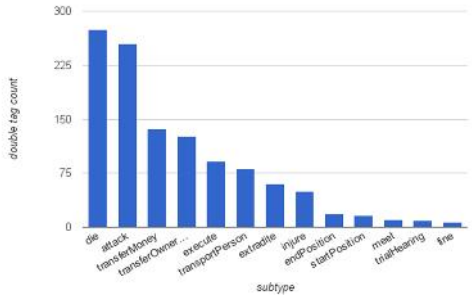


Figure 5: Distribution of event subtype of double tagging.

6 Annotation Challenges

One big challenge that annotators face is how to handle inferred events. Even though the annotation guidelines specifically instruct annotators not to tag any inferred events, whether an event mention is inferred not only depends on interpretation of the context, but also the meaning of the event triggers. Coupled with double tagging, this becomes a bigger issue. For example, in "*He then **trafficked** a large quatity of cocaine to the US.*", "trafficked" as defined as "buy or sell something illegally" is a trigger of Transaction.TransferOwnership event, but with the context of "to the US", it also indicates that the cocaine as been transported to the US. So it can also be a trigger for Movement.TransportArtifact event. One can argue that the Movement.TransportArtifact event is inferred, but it is still elusive to draw the boundary of inference.

As mentioned above, Contact event type poses a lot of difficulty and inference is one of the factors that contributes to this difficulty. The guidelines specify that regular speech verbs such as "say", "tell", and "speak" would be triggers for contact event types, but don't specify the other categories of reporting verbs, such as "argue", "advise", "order", or "testify", which involve more complicated interpretation. Some reporting verbs have not only the speech aspects, but also performative aspects, e.g., "testify", and "threaten". The question of whether to double tag such verbs (one for the contact type and one for the performative action) requires further discussion and clarification.

7 Conclusion and Future Work

In this paper, we described the annotation and evaluation of event nuggets and coreference in the TAC KBP evaluation. Annotated data has been distributed to DEFT and TAC KBP performers, and will be made available to the wider community as part of the Linguistic Data Consortium (LDC) catalog.

By analyzing the inter-annotator agreement between the two independent first pass annotations, we learned that some event types and subtypes are more difficult for annotators than others with respect to recall. We also learned that annotation consistency for event triggers has improved with the adoption of a minimal extent rule for event triggers in 2015 as compared with the maximal semantic extent unit rule in 2014's EN annotation, but inference is still a big challenge. Additionally, the contact event type still poses considerable difficulty. The addition of contact event attributes improved annotation consistency in 2015 as compared with that of 2014, but there is still room for improvement.

The detection and coreference of event nuggets provides an anchor for detecting event arguments and event-event relations. EN evaluation in TAC KBP 2015 attracted participation from 17 institutions, and NIST will continue to run an open evaluation of EN as part of the event track in TAC KBP 2016. The EN evaluation tasks will expand from English to multilingual, including Chinese and Spanish. Due to the scarcity of training data for some event types and subtypes, the set of event types and subtypes for the evaluation in 2016 will be reduced from 33 in 2014 evaluation and 38 in 2015 evaluation to 18 event types and subtypes.

Acknowledgments

This material is based on research sponsored by Air Force Research Laboratory and Defense Advanced Research Projects Agency under agreement number FA8750-13-2-0045. 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.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, Joe Ellis. 2014. A Comparison of the Events and Relations Across ACE, ERE, TAC-KBP, and FrameNet Annotation Standards. ACL 2014: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, June 22-27. *2nd Workshop on Events: Definition, Detection, Coreference, and Representation*.

Olga Babko-Malaya, Greg P. Milette, Michael K. Schneider, Sarah Scogin: Identifying Nuggets of Information in GALE Distillation Evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2012), Istanbul, May 21-27.

DARPA. 2012. Broad Agency Announcement: Deep Exploration and Filtering of Text (DEFT). Defense Advanced Research Projects Agency, DARPA-BAA-12-47.

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, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, Stephanie Strassel. 2015. Overview of Linguistic Resources for the TAC KBP 2015 Evaluations: Methodologies and Results. In *Proceedings of TAC KBP 2015 Workshop, National Institute of Standards and Technology*, Gaithersburg, Maryland, USA, November 16-17, 2015.

Marjorie Freedman. 2015. TAC: Event Argument and Linking Evaluation task Description. http://www.nist.gov/tac/2015/KBP/Event/Argument/guidlines/EventArgumentAndLinkingTaskDescription.v09.pdf

Marjorie Freedman and Ryan Gabbard. 2014. Overview of the Event Argument Evaluation. In *Proceedings of TAC KBP 2014 Workshop, National Institute of Standards and Technology*, Gaithersburg, Maryland, USA, November 17-18, 2014.

Linguistic Data Consortium. 2005. ACE (Automatic-Content Extraction) English Annotation Guidelines for Events Version 5.4.3.

Zhengzhong Liu, Teruko Mitamura, Eduard Hovy. 2015. Evaluation Algorithms for Event Nugget Detection: A pilot Study. *3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015).

Teruko Mitamura, Eduard Hovy. 2015. TAC KBP Event Detection and Coreference Tasks for English. http://cairo.lti.cs.cmu.edu/kbp/2015/event/Event_Mention_Detection_and_Coreference-2015-v1.1.pdf.

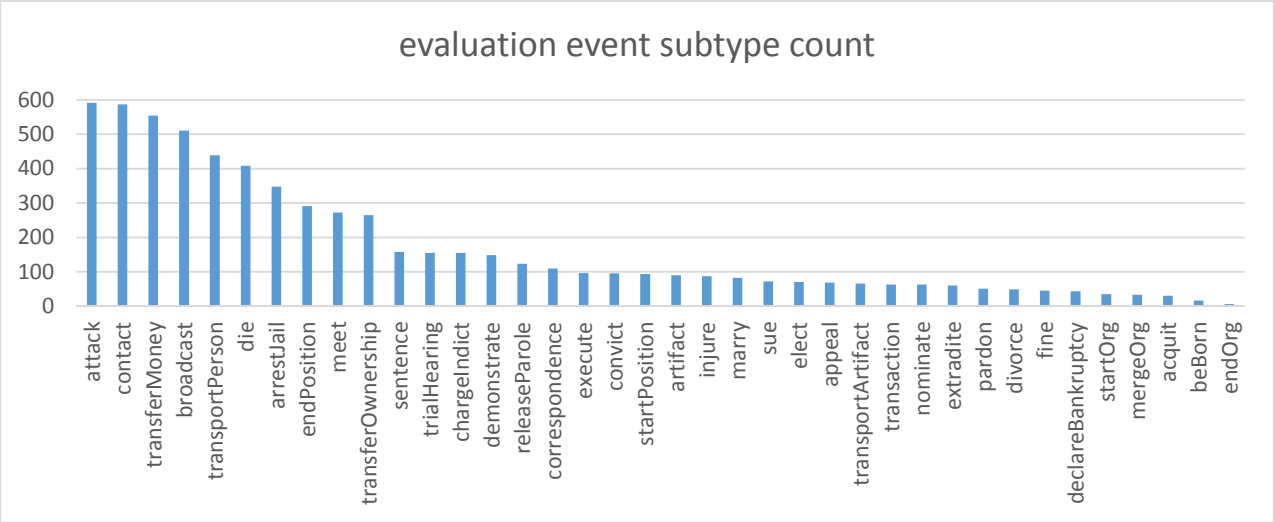
Teruko Mitamura, Yukari Yamakawa, Sue Holm, Zhiyi Song, Ann Bies, Seth Kulick, Stephanie Strassel. 2015. Event Nugget Annotation: Processes and Issues. *3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015).

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: the pyramid method. In Proceedings of the Human Language Technology Conference – North American chapter of the Association for Computational Linguistics annual meeting (NAACL-HLT 2004).

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. ACM Trans. Speech Lang. Process., 4(2):4.

Zhiyi Song, Ann Bies, Tom Riese, Justin Mott, Jonathan Wright, Seth Kulick, Neville Ryant, Stephanie Strassel, Xiaoyi Ma. 2015. From Light to Rich ERE: Annotation of Entities, Relations, and Events. *3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015).

Appendix



Event subtype distribution in Event Nugget 2015 Evaluation data