Events Detection, Coreference and Sequencing: What's next? Overview of the TAC KBP 2017 Event Track.

Teruko Mitamura **Zhengzhong Liu** **Eduard Hovy**
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213 USA
{teruko, liu, hovy}@cs.cmu.edu

Abstract

After two successful years of Event Nugget evaluation in the TAC KBP workshop, the third Event Nugget evaluation track for Knowledge Base Population(KBP) still attracts a lot of attention from the field. In addition to the traditional event nugget and coreference tasks, we introduce a new event sequencing task in English. The new task has brought more complex event relation reasoning to the current evaluations. In this paper we try to provide an overview on the task definition, data annotation, evaluation and trending research methods. We further discuss our efforts in creating the new event sequencing task and interesting research problems related to it.

1 Introduction

In TAC KBP 2017, we continued the traditional event nugget detection and coreference (hopper) tasks, the goal of which is to identify explicit mentions of events and within-document event coreference relations for three languages: English, Chinese and Spanish.

We follow the same setting of Event Detection and Hopper Coreference task from the previous year. The Event Detection task requires participants to detect events of 18 selected Event Types selected from the Rich ERE Annotation Guidelines: Events (v2.9.), as listed in Table 1. Also, the systems are required to identify the REALIS status (ACTUAL, GENERIC, OTHER) of the detected event nuggets. For the Event Nugget detection task, every instance of a mention of the relevant event types must be identified. If the same event is mentioned in several places in the document, participants must list them all.

The Event Coreference task requires participants to identify all coreference links among the event instances identified in a document, but not across document.

In this year we have introduced a new task: Event Sequencing. The task is based on the DEFT Event Sequence Pilot Evaluation Study, and it aims at detecting chronological relations of events focusing on a stereotypical sequence of (smaller) events that occur as part of a whole (larger) event.

2 Task Description

In this year there are two main tasks in the Event Nugget evaluation: each contains two sub tasks.

1. Task 1: Event Nugget Detection and Coreference
2. Task 2: Event Sequencing. This new task requires detecting **After** and **Subevent** relations, which is only available in English.

2.1 Task 1 Option 1: Event Nugget Detection

The **Event Nugget Detection task** aims to identify explicit mentions of relevant events in English, Chinese and Spanish texts. The inputs of this task are unannotated documents. The outputs are the identified event nugget span, each associated with event type and subtype labels, and REALIS tags.

Event Types and Subtypes: The participating systems must identify one of the event types and subtypes in Table 1. There are 7 event types and 18 event subtypes. For more details, see the Rich ERE

Annotation Guidelines: Events v.2.6 (Linguistic Data Consortium, 2015).

REALIS Identification: Event mentions must be assigned one of the following labels: ACTUAL (events that actually occurred); GENERIC (events that are not specific events with a (known or unknown) time and/or place); or OTHER (which includes failed events, future events, and conditional statements, and all other non-generic variations).

2.2 Task 1 Option 2: Event Coreference

The Event Coreference Task aims to identify the event nuggets and induce full event coreference links end-to-end. The event coreference task is to identify when two or more event nuggets refer to exactly the 'same' event. The annotation scheme for Event Coreference is called *Event Hoppers*. Event hoppers are relaxed notion of event coreference: each hopper contains events that "feel" coreferential to the annotator even if they do not meet the strict event identity (exactly same arguments between the nuggets)(Song et al., 2015).

The inputs of this task are unannotated documents. The outputs should contain event nuggets (with type/subtype and REALIS annotated), and event coreference links. Event type/subtypes are crucial for evaluation because they are required in aligning mentions (See §5).

2.3 Task 2: Event Sequencing

The new Event Sequence task aims to identify event sequence that occurs in a script (Schank and Abelson, 1977), which is a stereotypical sequence of events that occur as part of a whole (larger) event. Scripts help us understand a storyline composed of a series of events due to its well-known sequential progression.

In this task, we examine closely of two types of relations in scripts, which are shown in Figure 1. The AFTER links are between events following each other in a script, which is likely to be connected via their chronological order. Such ordering can be explicitly mentioned in text, or inferred from one's common knowledge. The SUBEVENT links are between a parent and a child event. The parent event is considered to be an event that include all its children. Chronologically, the parent event should cover the whole time period of the children events. For example:

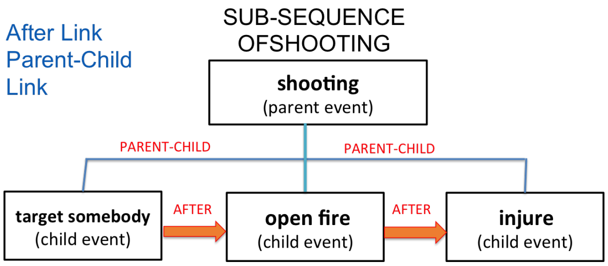


Figure 1: Event Sequence Link Structure

1. Thousands of people gathered **[E1, Movement.Transport-Person, ACTUAL]** in front of the city hall and chanted **[E2, Contact.Broadcast, ACTUAL]** the feminist slogan in Wednesday's protest **[E3, Conflict.Demonstrate, ACTUAL]**.

The event mention E3 ("protest") is the parent event for E1 ("gathered") and E2 ("chanted"). There is an "after" link between E1 and E2. Another sample event structure is shown in Figure 1.

3 Corpus

Event Detection and Coreference: The detection and coreference tasks are conducted in a trilingual setting. There are no newly annotated training corpus, but participants have access to the event nugget training and evaluation datasets from previous years. For the evaluation corpora, there are 167 documents for each language, half of the documents are newswire articles and the other half of the documents are texts from discussion forum. Table 2 list the the number of mentions of each type in the evaluation corpora. Comparing to previous years, the number of event mentions of each type is increased to better capture performance.

Event Sequencing: The training set for event sequencing are created by adding additional annotations on top of the TAC 2015 Event track training and testing data, which contains 360 documents. Since the event sequence task is annotated based on the events from the TAC 2015 Event Track, there will be 33 event types as listed in the TAC KBP 2015 Event guideline, instead of the 18 event types listed in this document. The test set for event sequencing are created on top of the TAC KBP 2016 Event Track data. Noe that the annotators are allowed to correct previous annotation errors during the annotation process.

Type	**Subtype**	**Type**	**Subtype**	**Type**	**Subtype**
Conflict	Attack	Transaction	Transfer Money	Manufacture	Artifact
Conflict	Demonstrate	Transaction	Transaction	Life	Injure
Contact	Meet	Transaction	Transfer Ownership	Life	Die
Contact	Correspondence	Movement	Transport.Artifact	Personnel	Start Position
Contact	Broadcast	Movement	Transport.Person	Personnel	End Position
Contact	Contact	Justice	Arrest-Jai	Personnel	Elect

Table 1: Event Types and Subtypes in TAC KBP Event Nugget 2016

Type	English	Chinese	Spanish	Type	English	Chinese	Spanish
conflict-attack	410	820	498	manufacture-artifact	141	131	67
conflict-demonstrate	209	101	166	movement-transportartifact	242	79	59
contact-broadcast	639	775	495	movement-transportperson	447	360	341
contact-contact	238	62	136	personnel-elect	152	176	141
contact-correspondence	198	25	101	personnel-endposition	186	108	97
contact-meet	200	126	115	personnel-startposition	116	78	106
justice-arrestjail	104	109	124	transaction-transaction	51	107	65
life-die	218	298	207	transaction-transfermoney	510	264	426
life-injure	53	93	93	transaction-transferownership	261	172	191

Table 2: Number of Nugget per Event Type in English

Most of the time, the annotators correct coreference links because they can be confused with or conflict with AFTER and SUBEVENT links.

4 Submissions and Schedule

Participating systems had about one week to process the evaluation documents for two tasks. Submissions must be fully automatic and no changes may be made to the system once the evaluation corpus has been downloaded. Up to three alternate system runs for each task may be submitted per team. Submitted runs should be ranked according to their expected overall score. Our timeline was as follows:

1. September 25 - October 2: Event Nugget Detection and Coreference evaluation
2. October 3 - October 10: Event Sequencing evaluation

5 Evaluation

5.1 Evaluating Event Detection and Coreference

We follow the evaluation metrics used in the last KBP Event Nugget Task (Liu et al., 2015; Mitamura et al., 2016). The event nugget evaluation scheme evaluates based on the best mapping between the system output and gold standard given the attributes being evaluated. Hence we have 4 metrics: (1) Span only: no attribute are considered other than span. (2) Type: consider the type attribute. (3) Realis: consider the Realis attribute and (4) All: consider all attributes.

For coreference evaluation, we used the mapping from (2) Type based mapping above. This is to deal with the Double Tagging problem where a single event nugget span may have two event types/subtypes. Coreference links normally link to one of the event types/subtypes only. Mapping with mention type may reduce the ambiguity in coreference evaluation. However, this also means coreference performance is highly influenced by the performance of event type classification. Also note that by using such mapping, we allow inexact mapping between system and gold standard mention span. We use the reference coreference scorer (Pradhan et al., 2014) to produce the coreference scores, and selected B^3, CEAF-E, MUC and BLANC. The systems are ranked based on the averaged of these 4 metrics.

5.2 Evaluating Event Sequencing

Similar to the Event Nugget Coreference, we first create a mapping between gold standard and system output. We then formulate two types of graphs from the relation Graph G: a closure of graph G^+ and a reduced graph $G-$. A closure graph is created by taking the transitive closure, and the reduced graph is a graph with redundant relations removed, where redundant relations are those which can be inferred through other relations.

We then calculate the performance of the "After" links and the parent-child links using the TempEval-3 evaluation methods (UzZaman and Allen, 2011; UzZaman et al., 2013). Here, the precision is calculated by first checking the number of the reduced system relations that can be verified from the reference closure graph, then divided by the total number of reduced system relations. Similarly, recall is calculated the other way around, by checking the portion of reduced reference links that can be verified by the system closure graph.

$$Precision = \frac{|Sys^-_{relation} \cap Ref^+_{relation}|}{|Sys^-_{relation}|}$$

$$Recall = \frac{|Ref^-_{relation} \cap Sys^+_{relation}|}{|Ref^-_{relation}|}$$

This evaluation metric utilizes the relation closure graph, which ensures that implicit relations are taken into account. For more detailed discussion of the evaluation metric, please refer to UzZaman and Allen (2011). In addition, UzZaman (2012) proposes an adjustment to this evaluation metric. The UzZaman and Allen(2011) metric give credits to all explicit relations, however, it does not give credits to systems that give more implicit relations. We include this adjusted metric in our scoring as well. For details of this adjusted metric, please refer to Section 6.1.7 of UzZaman (2012).

When constructing the temporal graph, we make the following assumption:

1. If E1 is linked to E2, any mentions sharing the spans with E1 should also link to E2 with the same relation type and direction.
2. The after link relations can be propagated through coreference links. If a co-referent link points to one nugget, all the nuggets in the cluster will be propagated.
3. There are no cycles allowed in the submitted file (validators are provided to check for cycles)

The first two assumptions allow links to be propagated. This propagation process is done during our annotation, according to the annotation guideline. However, it is not straightforward to do propagation automatically for submitted system results since it may create cycles in the temporal graph. As a result, we assume this step is done by the submitted systems.

Since this is the first time of event sequencing evaluation, we provide the gold standard event nuggets and coreference links to participants to focus on evaluating the performance of the sequencing task alone.

6 Results

In this section we provide an overview of the system performance on each language and each task. As discussed in §5, there are 4 different metrics for mention detection. To compare performance with these systems, one should focus on one of the evaluation metrics of interest. For example, event type will be more important for researchers who are interested in actual event content, while realis is more useful in determining the event status.

6.1 Overall Performance

6.1.1 English Nugget and Coreference

We summarize the performance of the participants on English Nugget Detection in Table 3 to 6. Each table lists the performance of each attribute group. Note that we only list the top performance from each team. English Nugget Detection results of all submissions are plotted in Figure 3 to 6.

The figures show that the top performing systems normally have a relatively balanced Precision-Recall trade off. In addition, we found that the most systems tend to have higher precision (blue lines) than recall (red lines). This trend is similar to what's observed in evaluation of year 2015 and 2016. Comparing to the gold standard, we found that most systems generally produce smaller number of event nuggets. The general low recall values indicate one major challenge in event nugget detection: how to predict event nuggets with new lexical items.

	Prec.	Recall	F1
srcb1	68.04	66.53	67.27
lvic-event1	64.89	55.71	59.95
UTD1	61.74	57.66	59.63
CLUZH1	57.34	61.09	59.16
TAMUNLP1	58.95	56.53	57.72
dsln-nlptt1	65.89	48.87	56.12
wip2	68.58	43.5	53.24
zy2	64.29	43.14	51.64
UI-CCG3	53.44	41.72	46.86
BUPT-PRIS1	67.95	32.74	44.19

Table 3: English Nugget Span Results

	Prec.	Recall	F1
srcb1	56.83	55.57	56.19
UTD1	52.16	48.71	50.37
lvic-event1	54.27	46.59	50.14
CLUZH1	47.1	50.18	48.6
dsln-nlptt1	57.02	42.29	48.56
wip2	60.98	38.68	47.33
TAMUNLP3	45.88	43.48	44.65
zy2	55.22	37.06	44.35
BUPT-PRIS1	58.92	28.39	38.31
UI-CCG3	37.46	29.24	32.85

Table 4: English Nugget Type Results

	Prec.	Recall	F1
CLUZH1	46.85	49.91	48.33
lvic-event1	51.39	44.12	47.48
srcb1	47.95	46.89	47.42
TAMUNLP1	43.38	41.6	42.47
dsln-nlptt1	49.86	36.98	42.47
UTD1	42.36	39.56	40.91
zy2	49.28	33.07	39.58
wip1	48.12	32.02	38.45
BUPT-PRIS1	46.36	22.34	30.15
UI-CCG3	30.3	23.65	26.57

Table 5: English Nugget Realis Results

	Prec.	Recall	F1
CLUZH1	38.51	41.03	39.73
lvic-event1	42.52	36.5	39.28
srcb1	39.69	38.81	39.24
dsln-nlptt1	43.22	32.05	36.81
UTD1	35.01	32.7	33.81
wip1	42.21	28.08	33.73
zy2	41.87	28.1	33.63
TAMUNLP3	33.35	31.6	32.45
BUPT-PRIS1	39.92	19.24	25.96
UI-CCG3	19.8	15.46	17.36

Table 6: English Nugget All Results

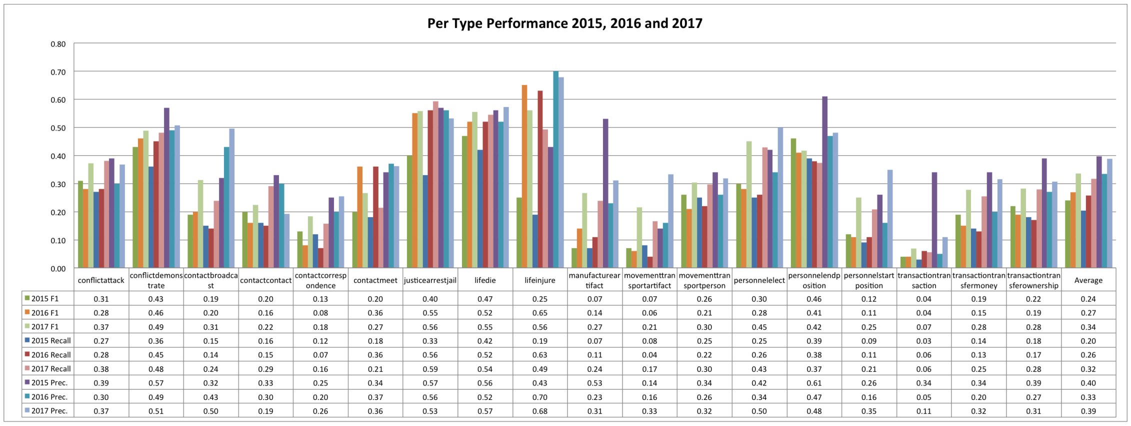
The nugget coreference results are summarized in Table 7. The top performance of coreference is higher than the last year systems by about 5 absolute F1 scores. This should be largely due to the better performance on event nugget detection.

	B^3	CeafE	MUC	BLANC	Aver.
srcb2	43.84	39.86	30.63	26.97	35.33
UTD2	39.88	35.73	33.79	26.06	33.87
TAMUNLP2	34.34	33.63	22.9	17.94	27.2
BUPT-PRIS1	28.66	28.64	19.3	13.56	22.54
UI-CCG3	24.98	23.36	12.57	8.96	17.47

Table 7: English Hopper Coreference Results

6.1.2 Nugget and Coreference performance on Chinese and Spanish

Due to the difficulties to build a system for other languages (Chinese and Spanish), only 3 teams participated in the Chinese task and 1 team participated in the Spanish task. We summarize the Chinese eval-



	conflictattack	conflictdemonstrate	contactbroadcast	contactcontact	contactcorrespondence	contactmeet	justicearrestjail	lifedie	lifeinjure	manufactureartifact	movementtransportartifact	movementtransportperson	personnelelect	personnelendposition	personnelstartposition	transactiontransaction	transactiontransfermoney	transactiontransferownership	Average
2015 F1	0.31	0.43	0.19	0.20	0.13	0.20	0.40	0.47	0.25	0.07	0.07	0.26	0.30	0.46	0.12	0.04	0.19	0.22	0.24
2016 F1	0.28	0.46	0.20	0.16	0.08	0.36	0.55	0.52	0.65	0.14	0.06	0.21	0.28	0.41	0.11	0.04	0.15	0.19	0.27
2017 F1	0.37	0.49	0.31	0.22	0.18	0.27	0.56	0.55	0.56	0.27	0.21	0.30	0.45	0.42	0.25	0.07	0.28	0.28	0.34
2015 Recall	0.27	0.36	0.15	0.16	0.12	0.18	0.33	0.42	0.19	0.07	0.08	0.25	0.25	0.39	0.09	0.03	0.14	0.18	0.20
2016 Recall	0.28	0.45	0.14	0.15	0.07	0.36	0.56	0.52	0.63	0.11	0.04	0.22	0.26	0.38	0.11	0.06	0.13	0.17	0.26
2017 Recall	0.38	0.48	0.24	0.29	0.16	0.21	0.59	0.54	0.49	0.24	0.17	0.30	0.43	0.37	0.21	0.06	0.25	0.28	0.32
2015 Prec.	0.39	0.57	0.32	0.33	0.25	0.34	0.57	0.56	0.43	0.53	0.14	0.34	0.42	0.61	0.26	0.34	0.34	0.39	0.40
2016 Prec.	0.30	0.49	0.43	0.30	0.20	0.37	0.56	0.52	0.70	0.23	0.16	0.26	0.34	0.47	0.16	0.05	0.20	0.27	0.33
2017 Prec.	0.37	0.51	0.50	0.19	0.26	0.36	0.53	0.57	0.68	0.31	0.33	0.32	0.50	0.48	0.35	0.11	0.32	0.31	0.39

Figure 2: Type Based Comparison 2015, 2016 and 2017

uation results in Table 9 and 10, and the Spanish results in Table 11 and 12. The best performance for Chinese and Spanish are lower than English, as expected. The F1 for Span is around 50 for both Chinese and Spanish. However, the best type based F1 in Chinese is 50.64, much higher than the 42.91 score for Spanish. It is encouraging to see the major performance improvement on Chinese. The relative larger gain in Chinese may due to the fact that there are relatively richer resources, such as parsers and datasets.

6.1.3 English Sequencing Performance

The Event Sequencing task is the new task being introduced this year, with only two participating teams. The best system run is submitted by KYOTOU, with a F1 score of 10.02 on After link detection (Precision: 7.52, Recall: 15.00), and 11.06 on Subevent link detection (Precision: 15.84, Recall: 8.49). These scores are low in comparison to the coreference tasks, and are similar to the results reported in the pilot study. The pilot study result is listed in Table 8.

7 System Approaches

As a general trend, many participants continue to use Neural Network (NN) based methods for event nugget detection. These year, both sequential NN models (RNN/LSTM) and Convolution Neural Networks are quite popular. Some systems employ hybrid learning models of both CNN and RNN. The input feature to the NN models mainly includes word embeddings and Part-of-Speech embeddings. Some systems have included features using n-hot representation. Traditional machine learning like SVM and CRF approaches are still popular among the participants. The top performing system in English ("srcb1") uses an ensemble result from CNN, LSTM and CRF models. Interestingly, the UTD team approach both the nugget typing and coreference with a multi-pass sieve approach and get the 2nd place in both English and Spanish. It may indicate that our current mention detection systems, equipped with Neural Networks, do not capture deeper semantics comparing to hand-crafted features. From another perspective, while it is claimed that the NN models are language independent, the performance difference is still significant between different languages. The performance is much better in English, a resource-rich language. It is definitely not straightforward to migrate a system easily across language.

The event sequencing systems are both NN based models, the KYOTOU system use a bi-directional GRU followed by a Multi-layer Perceptron (MLP); and the BOUN system use a MLP directly. However, there are no enough data to train good Neural Network Models, as indicated by the low performance.

8 Discussion

8.1 Challenges for Event Nugget Detection

In this section we discuss some main performance considerations for event nugget detection, based on the average value of all submission systems. Most values discussed here can be found in Figure 2.

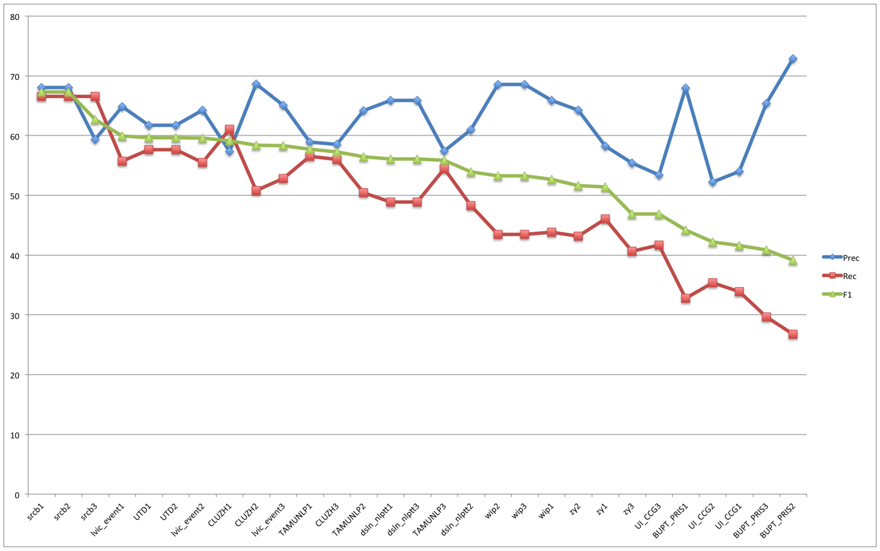


Figure 3: English Nugget Span Performance

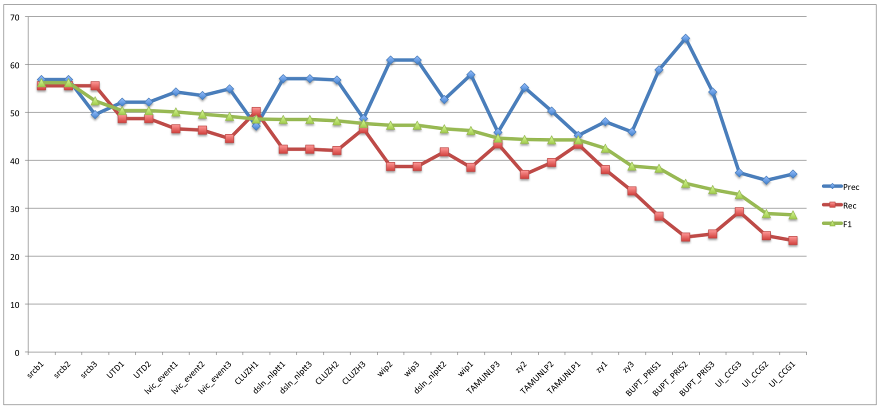


Figure 4: English Nugget Type Performance

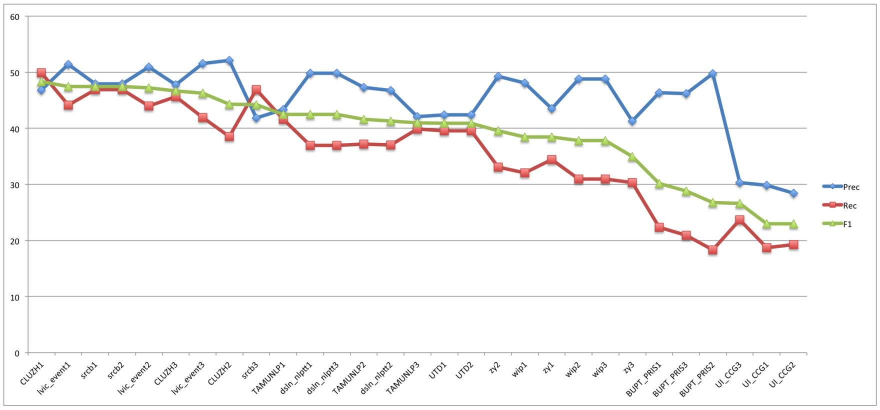


Figure 5: English Nugget Realis Performance

	Subevent			After			All		
System	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall
CMU1	-	-	-	18.5087	15.545	22.8688	-	-	-
CMU2	7.9083	9.4103	6.8198	17.5705	18.2877	16.9075	15.215	17.0509	13.736
OSU1	10.979	12.5749	9.7425	14.4474	14.6508	14.2495	13.9003	14.9867	12.9608
CMU3	7.3995	16.1215	4.8017	2.9427	8.2324	1.7916	4.4803	11.4833	2.7831
UC1	-	-	-	3.7678	36.4532	1.9865	-	-	-

Table 8: English Event Sequence Pilot Study Results

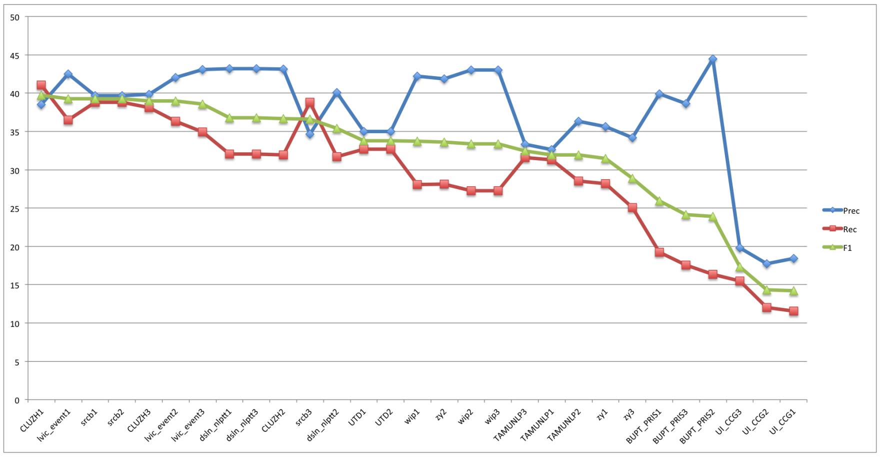


Figure 6: English Nugget All Performance

		Prec.	Recall	F1
Span	CLUZH1	67.76	45.92	54.74
	UTD1	52.69	53.02	52.85
	srcb2	47.48	46.76	47.12
Type	CLUZH1	62.69	42.48	50.64
	UTD1	46.61	46.91	46.76
	srcb2	42.47	41.82	42.14
Realis	CLUZH3	49.66	38.5	43.37
	UTD1	35.08	35.3	35.19
	srcb3	34.87	34.3	34.58
All	CLUZH3	45.76	35.48	39.97
	srcb3	31.77	31.25	31.51
	UTD1	31.07	31.27	31.17

Table 9: Chinese Nugget Evaluation Results

	B^3	CeafE	MUC	BLANC	Aver.
UTD1	34.18	32.22	27.07	18.57	28.01
srcb2	31.58	31.49	20.01	13.52	24.15

Table 10: Chinese Hopper Coreference Results

		Prec.	Recall	F1
Span	CLUZH2	60.93	42.64	50.17
	UI-CCG3	37.4	26.62	31.1
Type	CLUZH2	51.99	36.38	42.81
	UI-CCG3	27.96	19.9	23.25
Realis	CLUZH1	45.63	30.85	36.81
	UI-CCG3	21.17	15.07	17.6
All	CLUZH1	38.36	25.94	30.95
	UI-CCG3	15.26	10.86	12.69

Table 11: Spanish Nugget Evaluation Results

8.1.1 Difficult Event Types

Most participants suffer from a low recall problem in this evaluation. Interestingly, among all the event types, the event types that have low recall are almost identical comparing to last year. These event types include all subtypes under“contact”, “transaction” and “movement”. “manufacture.artifact” and “personnel.start-position” also have low recalls. A common characteristic of these event types is that they share very similar trigger and context words. For example, *Transaction-TransferOwnership* type refers to the events of transferring physical assets and *Transaction-Transaction* is used when it is unclear

	B^3	CeafE	MUC	BLANC	Aver.
UI-CCG1	9.9	10.39	3.89	2.04	6.55

Table 12: Spanish Hopper Coreference Results

whether the artifact in transaction is money or asset. Nuggets that trigger "contact" events are also very similar from the first glance. These event types are easily mis-classified into another. In addition, the trigger for contact events are sometimes very common words (e.g. say). Only a small portion of "say" are considered as communication events by the annotators. However, this is difficult to be captured by a learning system.

8.1.2 The Difficulties of Event Sequencing

The event sequencing task seem to be too challenging given the small number of data. In fact, the annotators are linking the nuggets based on common sense clues, which are not easily available to all the participant systems. Although the participant systems have tried to used external knowledges (e.g. VerbOcean and Conceptnet are used by the KYOTOU system), the performance of these system are still not comparable to more established tasks. Prior to this evaluation, successful computational models that captures script information are normally trained on large amount of data (Chambers, 2011). A more reasonable setup may be allowing systems to learn such knowledge from large scale domain data, and evaluate on the small set of annotated corpus.

9 Conclusion and Future Work

9.1 Evaluation Challenges

The coreference performance is difficult to be judged due to the low performance on nugget detection. Furthermore, event sequencing can also be viewed as a downstream task of event coreference. This year, we have released the gold standard coreference links to the participants. The performances of the systems are still low comparing to other established event tasks. Furthermore, annotating event relations can be a very expensive and error-prone process. This makes it difficult to produce a large amount and reliable datasets for training and testing.

9.2 The Difficulties of Richer Event Relations

As demonstrated by the evaluation result across these years, the main performance bottleneck for event nugget coreference is the performance of event nugget detection. One possible solution is to relax the strictness of event nugget matching algorithm: currently only nuggets that matches gold standard with the exact type and span will be considered.

In this year we have added the event sequencing relations, which relates to logical and temporal aspects of events. However, these tasks seem to be quite difficult both in terms of annotation and system development. The performance is still low, given the fact that systems have access to nuggets and coreference gold standard. It may be the case that the task is too difficult to be learned with such a small number of given data. In fact, the annotators are linking the nuggets based on common sense clues, which are not easily available to all the participant systems.

9.3 Embracing Learning with Large Scale Datasets

As shown in many research (Do et al., 2011; Huang et al., 2016; Peng et al., 2016), evaluation on limited annotated data may simply encourage systems to overfit the small number of selected event types. Furthermore, many difficult semantic interactions between events, such as argument relations and after relations, are difficult to be learned from small number of examples. It is a time to change the paradigm in the TAC-KBP event tracks, to allow training on massive data, and allow evaluations to be done on unrestricted domains.

References

Nathanael Chambers. 2011. *Inducing Event Schemas and Their Participants from Unlabeled Text.* Doctoral, Stanford University.

Quang Xuan Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R Voss, Jiawei Han, and Avirup Sil. 2016. Liberal Event Extraction and Event Schema Induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 258–268.

Linguistic Data Consortium. 2015. DEFT Rich ERE Annotation Guidelines: Events v.2.6. Technical report, Feb.

Zhengzhong Liu, Teruko Mitamura, and Eduard Hovy. 2015. Evaluation Algorithms for Event Nugget Detection : A Pilot Study. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 53–57.

Teruko Mitamura, Zhengzhong Liu, and Eduard Hovy. 2016. Overview of TAC KBP 2015 Event Nugget Track. In *TAC KBP 2015*, pages 1–31.

Haoruo Peng, Yangqi Song, and Dan Roth. 2016. Event Detection and Co-reference with Minimal Supervision. In *EMNLP 2016*.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (ii):30–35.

Roger C Schank and Robert P Abelson. 1977. *Scripts, Plans, Goals and Understanding*. 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 the 3rd Workshop on EVENTS at the NAACL-HLT*, pages 89–98.