Overview of TAC-KBP 2016 Event Nugget Track

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

Abstract

In this paper, we describe the second Event Nugget evaluation track for Knowledge Base Population(KBP) at TAC 2016. This year we extend the Event Nugget task to a trilingual setting: English, Chinese and Spanish. All the Event Nugget sub-tasks now require end-to-end processing from raw text. This task has attracted a lot of participation and intrigued interesting research problems. In this paper we try to provide an overview on the task definition, data annotation, evaluation and trending research methods. We further discuss issues related to the annotation process and the current restricted evaluation scope. With the lessons learned, we hope the next KBP Event Nugget task can incorporate more complex event relations on a larger scale.

1 Introduction

In the Event Nugget Track of TAC KBP 2016, our goal is to identify explicit mentions of events and event coreferences within the same text for three languages: English, Chinese and Spanish.

The Event Detection task is required to detect the selected Event Types and Subtypes taken from the Rich ERE Annotation Guidelines: Events (current version is v2.9.). Also, the task is to identify three REALIS ACTUAL, GENERIC, OTHER, which are described in the Rich ERE guidelines. For the Event Nuggest deteaction 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.

The eventual benefit of the Event Detection and Coreference will be to detect:

1. Cross-document, multi-lingual event coreference
2. Subevent structures or sub-sequence event linking

2 Task Description

There are two tasks in the Event Nugget evaluation for each three languages.

1. Event Nugget Detection
2. Event Nugget Detection and Coreference

This year we decided not to have the event coreference only task, since the matching baseline gives a relatively high score (Mitamura et al., 2016), and there was no specific reason to offer the event coreference only task this year.

2.1 Event Nugget Task 1

Event Nugget Detection task aims to identify explicit mentions of relevant events in English, Chinese and Spanish texts. The input of this task is unannotated documents. The output is event nugget tokens, event type and subtype labels, and REALIS information.

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. We have reduced the number of event types and subtypes from last year's evaluation, so that we can reduce the time for creating the training and evaluation sets. 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 Event Nugget Task 2

Event Nugget Detection and Coreference Task aims to identify full event coreference links and the event nugget detection task at the same time. Full event coreference is identified when two or more event nuggets refer to exactly the 'same' event. This notion is called *Event Hoppers* in the Rich ERE Annotation Guidelines. The full event coreference links do not include subevent relations.

The input of this task is unannotated documents. The output is event nuggets, event type and subtype labels, REALIS information, and event coreference links.

3 Corpus

Source data was provided by LDC. For English, about 200 annotated documents were provided prior to the evaluation as training set. For Chinese, a training set containing 200,000 words was provided. For Spanish, a training set containing 120,000 words was available.

For the formal evaluation, we planned to provide the minimum of 60 documents to the participants for each language. In the evaluation corpora, the minimum of 30 mentions of event types and subtypes were included. We evaluated both newswire articles and discussion forum text.

4 Submissions and Schedule

Participant systems had about two weeks to process the evaluation documents for two tasks for three languages. 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 20 - October 3: Event Nugget Detection evaluation
2. September 20 - October 3: Event Nugget Detection and Coreference evaluation

5 Evaluation

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 evaluate 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.

To evaluation coreference, 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 type/subtype. Coreference links normally link to one of the event type/subtype 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 the 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.

6 Results

In this section we provide an overview to the system performance on each language and each tasks. As discussed in §5, there will be 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

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

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 2 to 5. 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 most systems tend to have higher precision (blue lines) than recall (red lines). This trend is similar to what's observed in last year's evaluation[1]. Low recall values may indicate that some systems fail to generalize the observations of nuggets from the training data.

The nugget coreference results are summarized in Table 6. Since this year participants need to produce End-to-End coreference output, we only have 6 participating teams.

The final nugget and coreference result seems to drop significantly comparing to last year, even though the tasks are done in similar settings. The best mention type detection system this year is of score 46.99, while the best from 2015 had a F1 of 58.41. The top performer this year have an averaged coreference score of 30.08, while the best score of last year is 39.12. Part of this may be caused by the change in the evaluation types: many difficult and ambiguous evaluation types remains, such as Transaction.Transfer-Money and Transaction.Transfer-Ownership, and the Contact types. In fact, these types are quite popular in this year's test data, as shown in Figure 1.

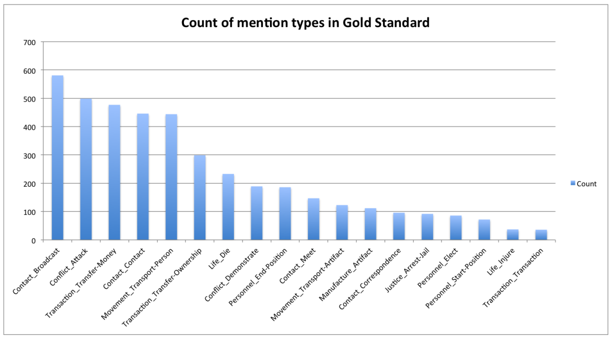


Figure 1: English Event Type Counts

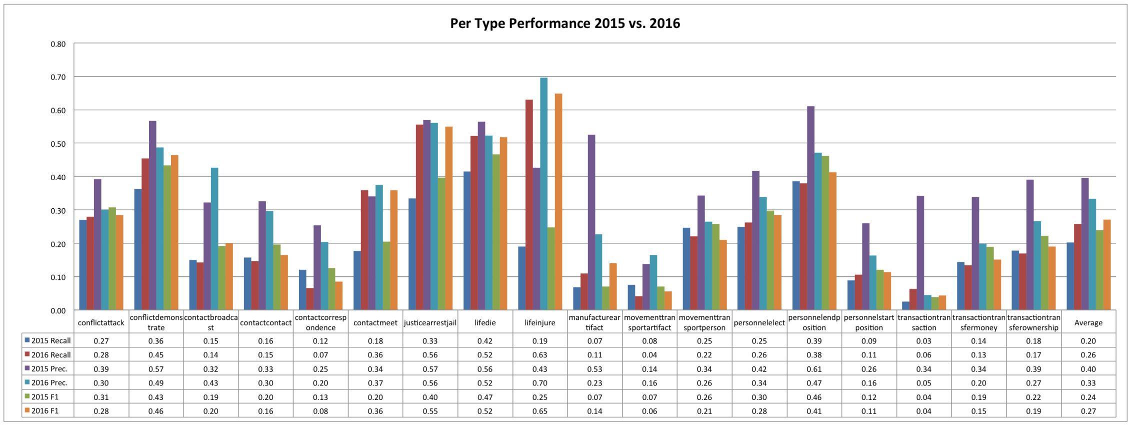
To show whether the actual performance is consistent across the two years, we take a closer look at the performance on the individual types. The per type performance evaluated in KBP 2016 is summarized in Figure 2. It can be seen that the averaged F1 performance (last column) is similar between two years. In fact, the F1 score on these types in year 2016 is slightly higher ($0.27 > 0.24$). Hence, we consider that the "performance drop" on nugget detection this year is only caused by the selection of event types to be evaluated. And since coreference depends highly on nugget detection, we also see a decrease in numbers on coreference performance.

Overall, the best nugget type F1 score is 46.99 and the best nugget realis score is 42.68 in English. There is still a long way to go to build a robust and accurate event nugget detection system.

6.1.2 Nugget and Coreference performance on Chinese and Spanish

Since this is the first time we introduce the task to Chinese and Spanish, only 4 teams participate in the

[1] we observe that the 2015 systems bias towards precision while the 2016 systems is more balanced (though also slightly biased towards precision). In last year's overview(Mitamura et al., 2016), we pointed out that the difference of type distribution between training and testing actually causing a low recall.



	conflictattack	conflictdemonstrate	contactbroadcast	contactcontact	contactcorrespondence	contactmeet	justicearrestjail	lifedie	lifeinjure	manufactureartifact	movementtransportartifact	movementtransportperson	personnelelect	personnelendposition	personnelstartposition	transactiontransaction	transactiontransfermoney	transactiontransferownership	Average
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
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
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

Figure 2: Type Based Comparison 2015 vs. 2016

	Prec.	Recall	F1
UTD1	55.36	53.85	54.59
NYU3	51.02	57.52	54.07
SoochowNLP3	58.11	45.17	50.83
LTI-CMU1	69.82	39.54	50.49
wip1	57.49	43.29	49.39
RPI-BLENDER1	51.48	46.11	48.65
aipheshd-t161	52.89	42.06	46.85
SYDNEY1	54.87	35.80	43.33
Washington1	50.19	35.02	41.25
HITS3	49.91	30.22	37.64
UMBC2	41.70	30.51	35.24
CMUML3	71.27	18.37	29.21
UI-CCG1	35.68	23.14	28.07
IHMC20161	6.66	5.02	5.72

Table 2: English Nugget Span Results

	Prec.	Recall	F1
UTD1	47.66	46.35	46.99
LTI-CMU1	61.69	34.94	44.61
wip1	51.76	38.98	44.47
NYU3	41.88	47.21	44.38
SoochowNLP3	49.92	38.81	43.67
RPI-BLENDER2	44.51	39.87	42.07
SYDNEY1	46.48	30.33	36.70
Washington1	42.15	29.41	34.65
aipheshd-t161	36.83	29.28	32.62
UMBC2	37.36	27.33	31.57
HITS3	41.79	25.30	31.52
CMUML3	60.44	15.58	24.77
UI-CCG2	25.81	18.53	21.57
IHMC20161	0.69	0.52	0.59

Table 3: English Nugget Type Results

Chinese task and 2 teams participate in the Spanish task. We summarize the Chinese evaluation results in Table 7 and 8, and the Spanish results in Table 9 and 10. The best performance on these two languages are both around 50 F1 for Span and around 40 F1 for Type detection, which are a couple points lower comparing to English. Besides from the fact that resources in languages other than English are normally scarce and less studied, there are also unique challenges in these languages. For example, Chinese is a very different language to English. Tokens in Chinese may be composed by several characters, sometimes 1 character is sufficient. On character tokens are more ambiguous and hence difficult to detect. In addition, we find that there are some annotation issues in the Chinese datasets, which is discussed in §8.3. In addition, unlike English and Spanish, datasets in Chinese are all of discussion forum genre; no newswire data annotated.

7 System Approaches

This year, many participants have introduced Neural Network based methods for event nugget detection. A popular choice of Neural Network architecture is LSTM and the problem is considered as a sequence labeling task. The NYU team uses a non-consecutive convolution neural network instead of a Recurrent Neural Network, which they consider the task to be a N-way classification problem on each token. They claim that such network architecture can incorporate position embeddings, which greatly improve the performance. Many other teams also feed other em-

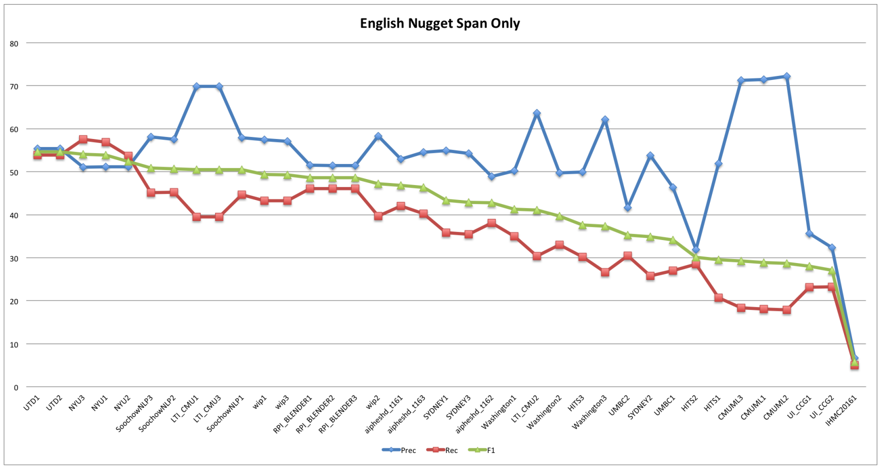


Figure 3: English Nugget Span Performance

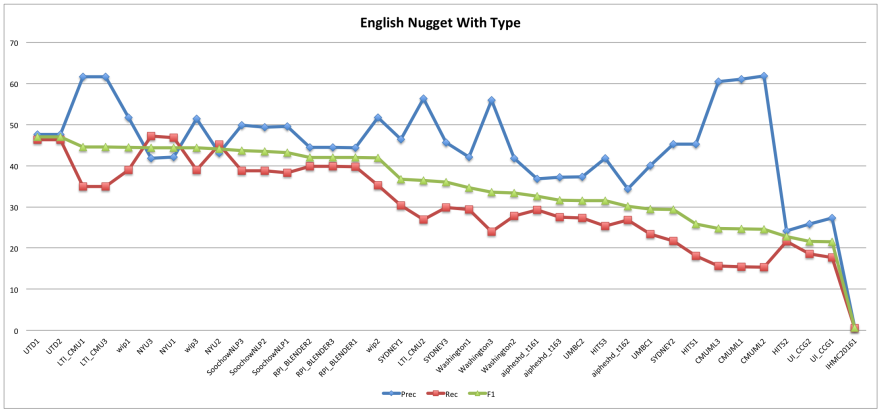


Figure 4: English Nugget Type Performance

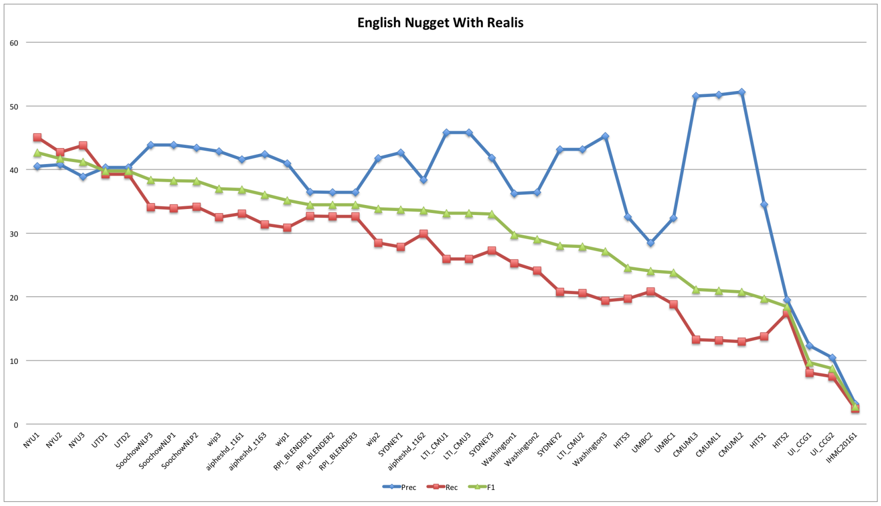


Figure 5: English Nugget Realis Performance

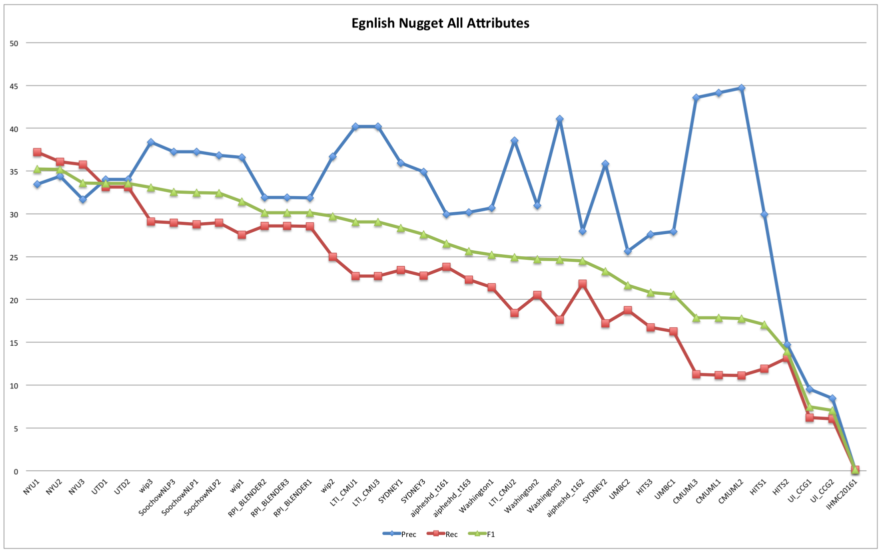


Figure 6: English Nugget All Performance

	Prec.	Recall	F1
NYU1	40.53	45.07	42.68
UTD1	40.34	39.23	39.78
SoochowNLP3	43.84	34.08	38.35
wip3	42.86	32.49	36.96
aipheshd-t161	41.57	33.06	36.83
RPI-BLENDER1	36.47	32.67	34.46
SYDNEY1	42.67	27.84	33.69
LTI-CMU1	45.78	25.93	33.11
Washington1	36.20	25.25	29.75
HITS3	32.56	19.72	24.56
UMBC2	28.45	20.81	24.04
CMUML3	51.54	13.29	21.13
HITS1	34.52	13.77	19.68
UI-CCG1	12.33	8.00	9.70
IHMC20161	3.20	2.40	2.75

Table 4: English Nugget Realis Results

beddings other than the word embedding into the system, including dependency relation embeddings and Part-of-Speech embeddings.

There are still some traditional feature based approaches for such problem. HITS improve the local, intra-sentential decoding to incorporate document level context. The LTI-CMU and CMUML teams also use submit systems using feature based methods.

	Prec.	Recall	F1
NYU1	33.47	37.21	35.24
UTD1	34.05	33.12	33.58
wip3	38.38	29.1	33.1
SoochowNLP3	37.26	28.97	32.59
RPI-BLENDER2	31.92	28.59	30.16
LTI-CMU1	40.19	22.76	29.06
SYDNEY1	35.93	23.45	28.38
aipheshd-t161	29.94	23.81	26.53
Washington1	30.71	21.42	25.24
UMBC2	25.64	18.76	21.67
HITS3	27.63	16.73	20.84
CMUML3	43.6	11.24	17.87
UI-CCG1	9.52	6.18	7.49
IHMC20161	0.13	0.1	0.11

Table 5: English Nugget All Results

8 Discussion

8.1 Challenges for Event Nugget Detection

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

8.1.1 Difficult Event Types

Most participants suffer from a low recall problem in this evaluation. We analyze the top misses from the systems. Among all the event types,

	B^3	CeafE	MUC	BLANC	Aver.
UTD2	37.49	34.21	26.37	22.25	30.08
LTI-CMU1	35.06	30.45	24.60	18.79	27.23
NYU1	34.62	33.33	22.01	18.31	27.07
RPI-BLENDER1	20.96	16.14	17.32	10.67	16.27
CMUML1	19.74	16.13	16.05	8.92	15.21
UI-CCG2	11.92	11.54	4.34	3.10	7.73

Table 6: English Hopper Coreference Results

		Prec.	Recall	F1
Span	LTI-CMU1	56.46	39.55	46.52
	UTD1	47.23	43.16	45.1
	LTI-CMU3	56.19	35.35	43.4
	UI-CCG1	28.34	39.61	33.04
	RPI-BLENDER1	62.46	18.48	28.52
Type	LTI-CMU1	50.72	35.53	41.79
	UTD1	41.9	38.29	40.01
	LTI-CMU3	49.7	31.26	38.38
	UI-CCG1	24.01	33.55	27.99
	RPI-BLENDER2	59.87	17.5	27.08
Realis	LTI-CMU1	42.7	29.92	35.18
	UTD1	35.27	32.23	33.68
	LTI-CMU3	43.11	27.12	33.29
	RPI-BLENDER2	48.46	14.16	21.92
	UI-CCG1	9.65	13.49	11.25
All	LTI-CMU1	38.91	27.26	32.06
	UTD1	31.76	29.02	30.33
	LTI-CMU3	38.54	24.25	29.77
	RPI-BLENDER2	46.69	13.65	21.12
	UI-CCG1	8.31	11.62	9.69

Table 7: Chinese Nugget Evaluation Results

Contact-Broadcast, Contact-Contact, Transaction-TransferMoney and *Transaction-TransferOwnership* are popular and have a recall value lower than average. These 4 event types contribute to around 50% of the total misses while they occupy 43% of the test data.

If we do not consider the popularity factor, the event types with the least recall values are *Transaction-Transaction, Manufacture-Artifact, Movement-TransportArtifact, Personnel-StartPosition*. All 4 types have a recall value lower than 10%. The majority of these misses are because systems do not produce any event nuggets. Less than 10% of the misses are because systems missing them with another event type.

	B^3	CeafE	MUC	BLANC	Aver.
UTD1	32.83	30.82	24.27	17.80	26.43
LTI-CMU1	30.83	26.95	24.07	17.67	24.88
RPI-BLENDER2	17.46	11.97	16.51	10.23	14.04
UI-CCG1	16.34	17.38	7.15	5.46	11.58

Table 8: Chinese Hopper Coreference Results

		Prec.	Recall	F1
Span	RPI-BLENDER1	48.11	51.31	49.66
	UI-CCG1	35.12	22.85	27.69
Type	RPI-BLENDER2	36.06	38.47	37.23
	UI-CCG1	26.01	16.92	20.50
Realis	RPI-BLENDER1	35.90	38.29	37.06
	UI-CCG1	11.03	7.17	8.69
All	RPI-BLENDER2	26.67	28.45	27.53
	UI-CCG1	8.43	5.48	6.64

Table 9: Spanish Nugget Evaluation Results

Some event types are very similar in their trigger words and sometimes even their context. For example, *Transaction-TransferOwnership* is the event of transferring physical assets and *Transaction-Transaction* is used when it is unclear whether the artifact in transaction is money or asset. These event types are easily mis-classified into another one.

For example, *Transaction-Transaction* type is mis-classified into *Transaction-TransferOwnership* for 114 times, while only correctly predicted 68 times. Another difficult case is *Movement-TransportArtifact*, which is mis-classified into *Movement-TransportPerson* for 467 times, while only correctly predicted 130 times. *Contact-Broadcast* is mis-classified into *Contact-Contact* for 1273 times, correctly predicted 3226 times.

Distinguishing such event types require detailed and deep semantic understanding. Normally one need to understand the argument of the event well to produce the correct prediction.

8.1.2 Chinese Single Character Nuggets

In discussion forum data, the language is normally less formal. In terms of Chinese event nuggets, We observe that a lot of them are of single tokens. In the Rich ERE Chinese training data, 17 are single character mentions in the top 20 most frequent event mention surfaces, This number is 6 in the ACE Chinese training data, which are newswire documents.

	B^3	CeafE	MUC	BLANC	Aver.
RPI-BLENDER1	22.05	18.56	19.04	12.43	18.02
UI-CCG1	11.25	10.5	6.76	3.24	7.94

Table 10: Spanish Hopper Coreference Results

A single character token in Chinese is normally quite ambiguous, for example, the token 打 can indicate an event of type ”attack” (打人) or ”call by phone” (打电话). Again, determining the event type here require analysis of the event argument.

8.2 Evaluation for Event Nugget Coreference

The main performance bottleneck for event nugget coreference is the performance of event nugget detection. Currently the low nugget detection performance make it very difficult for systems to produce reasonable results. Currently, we observe that simple type and realis based matching are still dominant features in predicting coreference links. We consider it may be a good idea to bring back the Coreference only task, where event nuggets are given beforehand, to show advancement in event coreference resolution.

8.3 Dataset Quality

We hypothesize that the Chinese datasets are not fully annotated. We take a closer look in the data and found a number of missed event nuggets. Here we list a couple examples:

(1) 支持香港同胞争取[Personnel.Elect 选举]与被[Personnel.Elect 选举]权！

(2) 司务长都是骑着二八去[TransferOwnership 买]菜去。

(3) 海豹行动是绝密，塔利班竟然可以预先得知？用个火箭就可以[Conflict.Attack打]下来，这个难度也实在是太高了吧。

In the above examples, we show several event nuggets. However, mentions annotated in red are not actually annotated in the Rich ERE datasets. Especially, in example 1, the first 选举 is annotated but the second one is not. Such inconsistencies happen a lot across the dataset. When training with such data, the classifier will likely to be quite conservative on making event nugget predictions. We conduct a very simple quantitative analysis by comparing the ACE 2005 Chinese annotation against the Rich ERE Chinese annotation. Table 11 and Table 12 summarize the top 5 double-character tokens annotated in ACE and RichERE. For the most popular event mentions, Rich ERE annotated only a smaller percentage comparing to ACE.

In addition, we find that the most popular event nuggets are mostly single character in the Rich ERE datasets, such as 打(170), 说(148), 死(131), 杀(118) . In fact, in top 20 most popular event nuggets of Rich ERE, there are 17 single-character nuggets, this number is only 6 in ACE. These single character tokens are more ambiguous comparing to a double character mention (for example, 打 can represent the action of ”calling someone” or ”attacking someone”, which corresponds to very different event type. This is because language in discuss forum posts are normally not formal. This actually challenges our event nugget systems to deal with deeper semantic problems.

Token	Annotated	Total	%
冲突	100	119	84.03%
访问	64	90	71.11%
受伤	53	59	89.83%
死亡	46	50	92.00%
前往	44	52	84.62%

Table 11: Top 5 double character mentions in ACE

Token	Annotated	Total	%
战争	96	223	43.05%
死亡	24	33	72.73%
暗杀	22	40	55.00%
入侵	18	22	81.82%
自杀	17	33	51.52%

Table 12: Top 5 double character mentions in ERE

9 Conclusion and Future Work

9.1 Evaluate Richer Event Relations

The current evaluation only focus on one type of link: coreference. However, events have temporal and spatial properties. There are many other relations between event nuggets. In next year’s event nugget track, we plan to introduce some other types of event relations: after (whether one event follows another), parent-child (whether one event include another event). We are planning to organizing a pilot-study before the actual evaluation next year.

9.2 Evaluation Challenges

As discussed in §8.2, the coreference performance problems are hidden by the low nugget performance, it may be a good idea to re-introduce the coreference only task.

In addition, annotating event nugget and coreference can be a very expensive and error-prone process. A cold-start style approach may be preferable. Systems can discover event nuggets with unsupervised or semi-supervised methods, given a small set of examples or definitions.

References

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.

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.