Overview of TAC-KBP 2015 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

This paper describes three TAC KBP Event Nugget tasks: (1) Event Nugget Detection, (2) Event Nugget Detection and Coreference, and (3) Event Nuggest Coreference. The evaluation corpus, prepared by LDC, consists of 202 documents from newswire and discussion forum. Participating systems detect event nuggets, event types and subtypes, and Realis values. For task 1, 38 runs were submitted by 14 teams; for task 2, 19 runs were submitted by 8 teams; for task 3, 16 runs were submitted by 6 teams. After the scoring algorithms and their results, we provide some analyses of these tasks.

1 Introduction

Analysis of events (recognition, coreference, linkage, etc.) of events in text is an important research area for deeper semantic understanding in natural language processing. Yet relatively few researchers have been working on this area. The complexity of definition and representation of events means there is no easy answer to the question of what events are, how we recognize them, and exactly how they relate to one another. In order to advance this research, we take a small and clear step towards the investigation of events.

In the Event Nugget Track of TAC KBP 2015, our goal is to identify explicit mentions of events and provide their coreferences within the same text. 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.

Within this Track, the Event Detection task focuses on detecting the Event Types and Subtypes as defined in the Rich ERE Annotation Guidelines: Events v2.6 (Linguistic Data Consortium, 2015; Language Technologies Institute - Carnegie Mellon University, 2015b). Also, the task includes assigning one of three REALIS values (ACTUAL, GENERIC, OTHER), which are also described in the Rich ERE guidelines. The data sources are provided by LDC. 158 annotated documents are provided prior to the evaluation as a training set. For the formal evaluation, 202 additional documents are given to participants. These documents include newswire articles and discussion forums.

The Event Coreference task requires participants to identify all coreference links among the event instances identified in a document. The intended benefit of the event detection and coreference is to detect and extract subevent configurations and produce an event ontology for future research.

2 Task Description

There are three tasks in the Event Nugget evaluation.

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

2.1 Event Nugget Task 1

Event Nugget Detection: This task aims to identify explicit mentions of relevant events in English text. Participating systems must identify all instances within each sentence, where relevance is defined by the event being one of the types/subtypes defined

Type	Subtype	Type	Subtype	Type	Subtype
Business	Start Org	Life	Divorce	Justice	Release-Parole
Business	End Org	Life	Injure	Justice	Trial-Hearing
Business	Declare Bankruptcy	Life	Die	Justice	Sentence
Business	Merge Org	Transaction	Transfer Ownership	Justice	Fine
Conflict	Attack	Transaction	Transfer Money	Justice	Charge-Indict
Conflict	Demonstrate	Transaction	Transaction	Justice	Sue
Contact	Meet	Movement	Transport.Person	Justice	Extradite
Contact	Correspondence	Movement	Transport.Artifact	Justice	Acquit
Contact	Broadcast	Personnel	Start Position	Justice	Convict
Contact	Contact	Personnel	End Position	Justice	Appeal
Manufacture	Artifact	Personnel	Nominate	Justice	Execute
Life	Be Born	Personnel	Elect	Justice	Pardon
Life	Marry	Justice	Arrest-Jail		

Table 1: Event Types and Subtypes

in the Rich ERE Annotation Guidelines (Table 1). In addition, systems must assign one of three REALIS values (ACTUAL, GENERIC, OTHER), which are also described in the Rich ERE guidelines and the TAC KBP Event Detection Annotation Guidelines v1.7 (Linguistic Data Consortium, 2015; Language Technologies Institute - Carnegie Mellon University, 2015b).

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: For purposes of this evaluation, an event must fall under one of the event types and subtypes in Table 1. 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).

Here are some example annotations of for the Event Nugget task:

(1) President Obama will `nominate` [realis: Other type: Personnel.Nominate] John Kerry for Secretary of State.

(2) He `carried out` the `assassination` [realis: Actual type: Life.Die] .

Event Nugget Identification: The definition of event nuggets generally follows the Rich ERE Annotation Guidelines. Each nugget is the actual string of words that indicates the mentioned event, and must correspond to the event type and subtype above. When a sentence mentions more than one event type both must be mentioned, e.g., in ‘’‘he shot the soldier `dead`”, both [Conflict.Attack] and [Life.Die] are events. We discuss how double-tagging spans are handled in §5.3.

2.2 Event Nugget Task 2

Event Nugget Detection and Coreference: In addition to the Event Nugget Detection task described above, this task also aims to identify full event coreference links 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.

2.3 Event Nugget Task 3

Event Nugget Coreference This task aims to identify full event coreference links, when the annotated event nuggets, event types and subtypes, and Realis

labels are given. The input of this task is the documents including this information. The output is event coreference relations for these given event nuggets.

3 Corpus

The evaluation corpus for this task contains 202 documents from two different types of documents: newswire and discussion forums. The original annotation is delivered by the Linguistic Data Consortium (LDC) in XML format.

3.1 Corpus Preprocessing

The annotations provided by LDC are based on character spans. Since character-based evaluation tends to assign higher weights to longer spans, we preprocess the corpus to provide a standard tokenized dataset using the Stanford CoreNLP toolkit (Manning et al., 2014). We also run a token boundary validator so that event mention spans do not stop in the middle of a mention. Our preprocessing step provides two new types of representation of the corpus: the Brat annotation tool format[1] and the Token Based Format (TBF). We believe that the Brat format make it easier for participants to view and even modify the annotations. For details of the conversion process, please refer to the scorer repository[2].

4 Submission Format

This section describes submission formats for all tasks. Our scorer accept the Token Based Format (TBF) as evaluation format. For each nugget detected, the system must output one line in a text file, using the following format for tab-separated fields:

- system-ID: unique ID assigned to each system run
- doc-ID: unique ID assigned to each source document
- mention ID: ID of the event nugget
- token ID list: list of IDs for the token(s) of the current mention
- mention-string: actual character string of event mention
- event-type: type.subtype from the hierarchy given above [3]
- Realis-value: one of ACTUAL, GENERIC, OTHER
- Confidence scores of event span: score between 0 and 1 inclusive (optional)
- Confidence scores of event type: score between 0 and 1 inclusive (optional)
- Confidence scores of Realis-value: score between 0 and 1 inclusive (optional)

Details of evaluation file formats are described in the Event Nugget Detection and Coreference Scoring v.27 document (2015a). If the system chooses not to provide the confidence scores, then the last three fields are empty.

Coreference decisions are attached after listing all nuggets in a document. Each coreference cluster is also represented in one tab-separated line, using the following columns:

- Relation name: this should always be @Coreference
- Relation Id: This is for bookkeeping purposes, which will not be read by the scorer. The relation id used in the gold standard files will be in form of "R[id]" (e.g., R3)
- Mentions Id list: list of event mentions in this coreference cluster, separated by comma. In terms of coreference, the ordering of event mentions does not matter.

In addition, special headers and footers are used to mark boundaries of documents.

5 Scoring

An automated scorer that we have created reads the output of event mention detection systems and compares them to the gold standard. In general, for event nugget detection, systems are scored using the F-1 score that balances Precision and Recall compared to the gold standard. For event nugget coreference, systems are scored using the evaluation metrics used in CoNLL shared tasks.

[1] brat.nlplab.org

[2] http://hunterhector.github.io/EvmEval/

[3] Upon the request of some participants, the type.subtype format is normalized before scoring: punctuation marks are removed and all characters are lower-cased.

The input and output of the scorer are:

Input:

1. Gold standard annotation for a text, in evaluation file format (tbf)
2. System output annotation for the same text, in evaluation file format (tbf)
3. Standard token table provided to participants.

Output:

1. System score report for event nugget detection and coreference.
2. Optional system gold difference report[4].

5.1 Event Nugget Detection Evaluation

We used a slightly updated version of the attribute-aware scoring metric described in Liu et al. (2015).

Mention Mapping: In order to evaluate mention attributes (such as REALIS labels, event types, etc.), the evaluation algorithm first needs to decide which system mention corresponds to a gold standard mention. We refer to this step as mention mapping. The input of our mention-mapping algorithm is the pairwise scores between all gold standard vs. system mention pairs, measured using a token-based Dice score[5]. Algorithm 1 shows our mapping algorithm to compute the mapping in one document.

Computing the F score: Given the mapping, we then compute the True Positive(TP) values using algorithm 2 for each attribute configuration. We can choose the set $\mathcal{A}$ to contain the desired attributes we would like to evaluate on. Note that when we choose $\mathcal{A}$ to be the empty set, we will reduce to the span-only scoring. In our implementation, we iterate all possible attribute combinations and report all the scores (i.e., span only, mention type only, realis status only, and all).

With the TP value for each attribute configurations, we compute the false positives FP as $N_S - TP$, and then the Precision and Recall calculations are:

$$P = \frac{TP}{N_S}; R = \frac{TP}{N_G}$$

where N_S and N_G are the number of system mentions and gold standard mentions respectively. We then compute F1 by taking the harmonic average of P and R.

Algorithm 1 Compute a mapping between system and gold standard mentions with attributes

Input: A list L of scores $Dice(T_G, T_S)$ for all pairs of G, S in the document

Input: The set $\mathcal{A}$ indexing the attributes that will be evaluated for all mentions

$M \leftarrow \varnothing$; $U_s \leftarrow \varnothing$; $U_g \leftarrow \varnothing$;
while $L \neq \varnothing$ **do**
  $G_m, S_n \leftarrow \arg\max_{(G,S)\in L} Dice(T_G, T_S)$
  $L \leftarrow L - \{Dice(T_{G_m}, T_{S_n})\}$
  if $S_n \notin U_s$ **and** $G_m \notin U_g$ **and** $Dice(T_{G_m}, T_{S_n}) > 0$ **then**
    if $\mathcal{A}_{S_n} = \mathcal{A}_{G_m}$ **then**
      $M_{G_m} \leftarrow (S_n, Dice(T_{G_m}, T_{S_n}))$
      $U_s \leftarrow U_s \cup \{S_n\}$
      $U_g \leftarrow U_g \cup \{G_m\}$

Output: The mapping M

Algorithm 2 Compute True Positive from mapping

Input: The set of gold standard mentions $\mathcal{G}$;

Input: The mapping M indexed by gold standard mentions;

$TP \leftarrow 0$
for $G \in \mathcal{G}$ **do**
  $(S, Dice) \leftarrow M_G$
  $TP \leftarrow TP + Dice$

Output: TP

5.2 Coreference Evaluation

To evalulate event mention coreference, We follow the practice of CoNLL shared tasks on Entity Coreference. We apply all 4 popular metrics used by the community (MUC (Chinchor, 1992), B^3 (Bagga and Baldwin, 1998), and $CEAF-E$ (Luo, 2005)) and take the average of their scores to provide a unified score. In addition we also include the $BLANC$ (Recasens and Hovy, 2011) measure in the final average[6]. Details of evaluation scoring appear in the Event Nugget Detection and Coreference Scoring v.27 document(Language Technologies Institute - Carnegie

[4] A simple visualizer is hosted in the scorer repository that can render a web based difference view using this report.

[5] The Dice coefficient between the two token sets, which is the same as the F-1 score.

[6] BLANC was omitted in earlier CoNLL shared tasks because its scorer was not ready

Mellon University, 2015a)[7]. The scorer is currently maintained at a public repository[8].

Coreference metrics are complicated and are difficult to implement in practice. We therefore convert our coreference decisions to CoNLL format and feed them to the standard reference scorer (Pradhan et al., 2014). Our scorer output also includes raw output from the reference scorer. The conversion is done by taking the gold-system-alignment result from the mention type detection mapping[9].

In addition, we consider only exact-aligned mention pairs following the convention in entity coreference practice (e.g., mention "attack" will not be considered to be aligned with""conduct attack"). We leave partial mapping in the evaluation to future work.

5.3 Double(Multi) Tagging

An important difference between the current event coreference annotations with standard entity coreference is the existence of double tagging in our data. Here, a particular event mention can sometimes be annotated to refer tomore than one different instances or events types, for example:

(3) the murder of John on Tuesday and Bill on Wednesday.

1. **murder**, argument=John, time=Tuesday
2. **murder**, argument=Bill, time=Wednesday

(4) the murder of John and Bill

1. Conflit.Attack, **murder**
2. Life.Die, **murder**

The phenomenon of multi-tagging is caused by the fact that multiple event instances can be triggered by the exact same span. This creates some difficulty for scoring. For instance, the CoNLL scorer distinguishes mentions solely based on the mention span so that the coreference score will be different when the scorer make different decisions on aligning the mentions. One potential solution is to find the best possible scores for all possible mappings, which require enumerating all possible alignments between the system result and gold standard. We find this solution computationally infeasible, since there may sometimes be as many as 50 instances of double-tagging in a single document.

Our current solution is to perform greedy alignment in mention detection: each time we pick the best available mention alignment for each gold standard mention (see the mention scoring algorithm above for details). We break ties arbitrarily, following the order they appear in the result file.

For coreference scoring, we use the alignment from the mention-type mapping stage of event nugget detection. By also enforcing mention-type mapping, we try to reduce the ambiguity of double tagging and its effect on coreference. Currently we are investigating methods to eliminate such ambiguity but we did not have enough time to implement them for this evaluation.

5.4 Validation

Two validation measures are implemented in the scorer, and a stand-alone output validator is also provided. The purpose of the validation is to discover obvious format errors in submission and reject improper results that may alter a system's real performance. Besides standard format check, the following special validations have been employed:

1. Mentions in the same cluster cannot have the exact same span.
2. Different clusters cannot have mentions in common.
3. Mentions that appear in clusters should also appear in the mention list.
4. Mentions cannot have tokens not included in the token list provided.

6 Submissions and Schedule

Participant systems have about one week to process the evaluation documents. 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

[7] http://cairo.lti.cs.cmu.edu/kbp/2015/event/Event-Mention-Detection-scoring-v27.pdf

[8] http://hunterhector.github.io/EvmEval/

[9] The mapping from Alg. 1 where $\mathcal{A}$ only contains mention type.

submitted per team. Submitted runs should be ranked according to their expected overall score.

Our timeline was as follows:

1. September 8–21: Event Nugget Detection evaluation
2. September 8–21: Event Nugget Detection and Coreference evaluation
3. September 21–29: Event Nugget Coreference evaluation

7 Results

Seventeen teams submitted their runs to one or more Event Nugget tasks. Official scores were computed using the gold standard annotations in TAC KBP 2015 Event Nugget and Event Coreference Linking (LDC2015R26) and using the official KBP scorer. In follow-up investigations we found that the official scorer favors recall due to a particular way of mapping predicted to gold standard nuggets. To balance with precision, We subsequently modified the scorer and compute a set of new scores. The original official scorer is version 1.6, and the modified scorer is version 1.7[10]. This change affected the rank-ordering of only two or three systems, as shown in Table 6 (official) and Table 7 (updated).

Starting in 2016, the new scoring measure will be used. In this paper, We present both sets of results. We denote the KBP official results as **official**, and the results produced by the modified scorer as **update**. We encourage that future comparisons should be performed against the **updated** results.

7.1 Task 1: Event Nugget Detection Results

For this task, 38 systems were submitted by 14 teams. We report micro-average F1 for 4 attributes: span only (Table 2); REALIS (Table 3); type (Table 4); all attributes (Table 5). For each metric, we report only one run for each team (the one with the highest F1 score). Since different systems have different strengths on different attributes, their F1 rankings in these tables differ.

7.2 Task 2: Event Nugget Detection and Coreference Results

For this task, 19 runs were submitted by 8 teams. We report the official results in Table 6 and the update results in Table 7. Taking into account the small ranking change from the updated scoring algorithm, we report the highest averaged coreference score from each team. In the updated results (Table 7), we also report the system's performance on detecting the nuggets for future reference. From the results, we observe that one major constraint on coreference performance comes from detecting the correct event mentions.

[10]Both are available at https://github.com/hunterhector/EvmEval.

	Precision	Recall	F1
Team13	74.86	57.92	65.31
Team9	81.99	52.02	63.66
Team6	82.46	50.3	62.49
Team5	78.59	49.53	60.77
Team16	79.4	48.61	60.3
Team12	82.22	46.99	59.8
Team10	65.43	54.86	59.68
Team3	66	50.72	57.36
Team11	59.08	52.11	55.38
Team1	45.8	58.5	51.38
Team4	51.48	41.62	46.03
Team2	89.5	24.55	38.53
Team17	82.39	21.82	34.5
Team14	40.76	28.88	33.81

Table 2: Best event nugget span detection results for each team (Micro-Average)

7.3 Task 3: Event Nugget Coreference Results

For this task, 16 runs were submitted by 6 teams. We report the submission scores[11] in Table 8 (official) and Table 9 (updated). We also added 2 baseline systems as described in §7.4.

7.4 Baselines

We created two simple baselines for the coreference-only task. The **Singleton baseline** (Row S in Table 9) is generated by placing each individual mention into its own cluster. The **Matching baseline** is generated simply by considering all mentions with the same mention type and REALIS status to be coreferent[12].

From the table one sees that the Singleton (S) baseline is strong in several metrics, notably very high

[11]One submission is omitted due to formatting errors.

[12]One exception is that the scorer explicitly disallows two mentions with the same type and same span to be in the same cluster, for such cases, we simply retain mentions that appear later in the file.

	Precision	Recall	F1
Team9	75.23	47.74	58.41
Team5	73.95	46.61	57.18
Team6	73.68	44.94	55.83
Team13	73.73	44.57	55.56
Team16	71.06	43.5	53.97
Team10	66.77	42.53	51.97
Team12	67.95	38.83	49.42
Team3	55.42	42.59	48.16
Team11	45.59	40.21	42.73
Team4	46.49	37.59	41.57
Team1	31.35	40.05	35.17
Team2	78.55	22.24	34.67
Team17	77.85	20.62	32.6
Team14	32.46	23.00	26.93

Table 3: Best event nugget type detection results for each team (Micro-Average)

	Precision	Recall	F1
Team13	56.35	43.6	49.16
Team9	62.73	39.8	48.7
Team6	62.09	37.87	47.05
Team10	49.93	41.86	45.54
Team16	57.79	35.38	43.89
Team12	58.94	33.68	42.87
Team5	52.18	32.89	40.35
Team11	40.86	36.05	38.3
Team1	33.37	42.63	37.44
Team3	42.62	32.75	37.04
Team4	34.91	28.22	31.21
Team2	65.41	17.94	28.16
Team17	58.43	15.47	24.47
Team14	21.81	15.45	18.09

Table 4: Best event nugget REALIS detection results for each team (Micro-Average)

in terms of B^3. This is probably caused by the large number of singletons in the dataset. The MUC scorer give zero scores to singletons, since it credits only links[13]. As our final score does not include Macro-averages at the document level, this issue does not affect system scoring.

The performance of the Matching (M) baseline is very competitive: it ranks 6th over all the submissions. This shows that mention types and REALIS status both contain important information. Some event mentions are very difficult to resolve without type information, and sometimes type information is the determining factor. For example, in 5 to understand the relation of the two mentions `death` and `succumbed`, one needs to preform complex analysis such as complicated entity coreference on *Jack Layton* and *the former NDP leader*. The lexical senses of the two mentions provide very little information. However, by knowing that both mentions are of type Life.Die, we increase confidence for making the coreference decision. In fact, this pair of mentions is coreferent. In this task, thanks to the fine-grained ontology of event types and the domain-specific focus of the documents, the chance of coreference when the types of two mention matches is very high.

(5) Prime Minister Stephen Harper could at any time call a by-election in the riding of Toronto-Danforth which was left vacant by the `death` of Jack Layton. He must do so by Feb. 22, six months after the former NDP leader `succumbed` to cancer.

[13]https://github.com/conll/reference-coreference-scorers/issues/2.

	Precision	Recall	F1
Team9	56.98	36.16	44.24
Team6	55.12	33.62	41.77
Team13	47.04	36.39	41.04
Team16	52.12	31.9	39.58
Team10	43.12	36.16	39.33
Team5	49.22	31.02	38.06
Team12	49.88	28.50	36.28
Team3	57.83	23.36	33.27
Team11	31.65	27.92	29.67
Team4	31.71	25.63	28.35
Team2	57.88	16.39	25.54
Team1	22.11	28.25	24.81
Team17	55.68	14.75	23.32
Team14	16.74	11.86	13.89

Table 5: Best event nugget (all attributes) detection results for each team (Micro-Average)

The performance of participant are summarized in Table 8 (official) and Table 9 (updated), where we provide the highest averaged score for each team. The difference between the official and updated result is very small, and the ranking does not change. This

	Coreference score
Team9	63.23
Team5	62.95
Team12	60.33
Team8	55.67
Team17	53.57
Team15	52.48
Team1	26.33
Team14	17.80

Table 6: Official event nugget detection and coreference task results (Micro-Average of 4 metrics)

	Plain	Type	Realis	Type & Realis	Coref score
Team5	60.77	57.18	40.35	38.06	39.12
Team9	62.13	57.41	47.85	43.73	37.23
Team15	64.56	57.45	45.21	39.67	32.36
Team12	59.8	49.42	42.87	36.28	31.39
Team8	46.67	39.47	32.13	27.44	21.71
Team1	51.38	35.17	37.44	24.81	14.82
Team17	34.5	32.6	24.47	23.32	13.87
Team14	33.81	26.93	18.09	13.89	6.36

Table 7: Updated event nugget detection and coreference results (Micro-Average of 4 metrics)

	B^3	CEAF-E	MUC	BLANC	Avg.
Team5	82.85	74.66	68.5	77.62	75.69
Team12	83.75	75.81	63.78	73.99	74.28
Team6	82.27	75.15	60.93	71.57	72.6
Team13	82.18	75.45	51.42	68.88	70.02
Team9	81.6	75.43	51.4	68.85	69.94
Team7	84.72	77.42	0.00	48.75	56.88
A	80.83	73.55	52.01	66.67	68.72
S	78.1	68.98	0.00	48.88	52.01
M	78.40	65.82	69.83	76.29	71.94

Table 8: Official coreference-only results (A is the averaged score of each column; S is the singleton baseline; M is the simple Mention Type + Realis Match baseline.)

is because when participants start with gold standard event nuggets the influence of event nugget mapping algorithm is very small.

8 Discussion

In this section we present some simple corpus statistics. We have found that the number of certain mention types has changed significantly. In addition, we also find some differences of distribution in document length between the two datasets. We also present an analysis on the two different document genres.

	B^3	CEAF-E	MUC	BLANC	Avg.
Team5	82.29	74.12	68.08	76.91	75.35
Team12	83.09	75.36	63.16	73.2	73.7
Team6	82.27	75.14	60.9	71.56	72.47
Team13	82.18	75.45	51.45	68.88	69.49
Team9	81.6	75.42	51.37	68.84	69.31
Team7	52.92	57.41	0.00	21.36	32.92
A	80.83	73.55	52.01	66.67	68.72
S	78.1	68.98	0.00	48.88	52.01
M	78.40	65.82	69.83	76.29	71.94

Table 9: Updated coreference-only results (A is the averaged score of each column; S is the singleton baseline; M is the simple Mention Type + Realis Match baseline.)

8.1 Comparing Training and Testing Documents

Some participants have observed a performance drop on the event nugget detection task. Here we present a small corpus analysis to see if there are any high level differences between the training and testing datasets. We summarize the main figures in Table 10.

Statistics	**Training**	**Test**
# Docs	158	202
# Mentions	6538	6438
# Clusters	1154	1050
# Tokens	139444	98414
# Singleton	2185	3075
Aver. Mention Per Doc	41.38	31.88
Aver. Token Per Doc	882.56	487.20
# Token/ # Mention	21.33	15.29
Double tagged Mentions	323	575
Aver. Cluster Size	3.77	3.20

Table 10: Corpus comparison of training and testing datasets

We have also observed that the type distributions differ between training and test documents. Figure 1 shows the top 15 types in the training corpus and their counts in the training and testing corpus respectively. The training set has more than 800 Conflict.Attack mentions, while the test set has less than 600. Another notable difference is that the number of Con-

tact.Contact event mentions is around 600, almost double its count in the training set. The numbers of Justice.Pardon and Justice.Convict mentions also significantly decrease in the test set.

The two sets of documents exhibit have some differences in terms of discourse length. The average length (in terms of tokens) of training documents is 882.56, almost double the average length of test documents (487.20). The ratio token/mention shows that event mentions are sparser in the test set. Such differences may influence both mention type detection and coreference. For example, in a per-token annotation model, the chance that a tokens is a mention is smaller in the test set. In a coreference system, this may affect features that depend on discourse distance.

The average size of clusters differs by 0.57 across the two datasets. Such dataset differences might be introduced by some large documents. In fact, the largest cluster in the training set contains more than 70 mentions while the largest cluster in the test set contains only 18 mentions. The longest document in the training set contains 5616 tokens, while the longest document in the test set contains only 1127 tokens.

8.2 Genre Differences

We conducted a similar analysis on different genres on the training data. The results are summarized in Table 11. We summarize several observation from this table.

1. Forum documents are significantly longer than newswire documents and contain more mentions.

2. Event mentions in newswire are more dense (token/mention ratio is 13.64) compareds to forum data (25.28). This may affect performance of both detection and coreference, especially for systems that use features related to discourse distance.

3. Mentions tend to form larger clusters in forum documents than in newswire data. The average cluster size in forum is 4.03 compared to 3.16 in newswire. In addition, the number of singletons in forum data is smaller even though there are many more mentions.

Stat.	**News**	**Forum**
# Docs	81	77
# Mentions	2219	4319
# Clusters	350	804
# Tokens	30257	109187
# Singleton	1112	1073
Aver. Mention Per Doc	27.48	56.09
Aver. Token Per Doc	373.54	1418.01
# Token/ # Mention	13.64	25.28
Aver. Cluster Size	3.16	4.03

Table 11: Corpus comparison of Newswire and Forum on training dataset

9 Conclusion

The KBP Event Nugget task has attracted many participants, which shows that the community is interested in the research of events. However, the evaluation results have shown this task is very difficult. The best event mention detection system F1 score is lower than 0.50. In addition, the best coreference system is still very close to the simple type and realis matching baseline. Deeper understanding and analysis of event mentions is needed to change this situation.

It is always interesting to learn lessons from a similar research field. The tasks of event mention detection and coreference share many similarities with entity coreference. However, there are also some important differences. For instance, an event is normally comprised of a complex structure. To fully resolve an event mention, one may need to resolve all its arguments.

Another interesting comparison of these two areas are the differences on their ontology granularity. Our approach to event mention annotation uses a fine-grained type definition (which might be due to the nature of event semantics). Evaluation results have shown that such annotation schemes have moved some of the challenges of mention coreference towards event nugget and type detection. We hypothesize that better modeling on mention detection, especially modeling the interaction of mention detection and coreference, will be important future research steps.

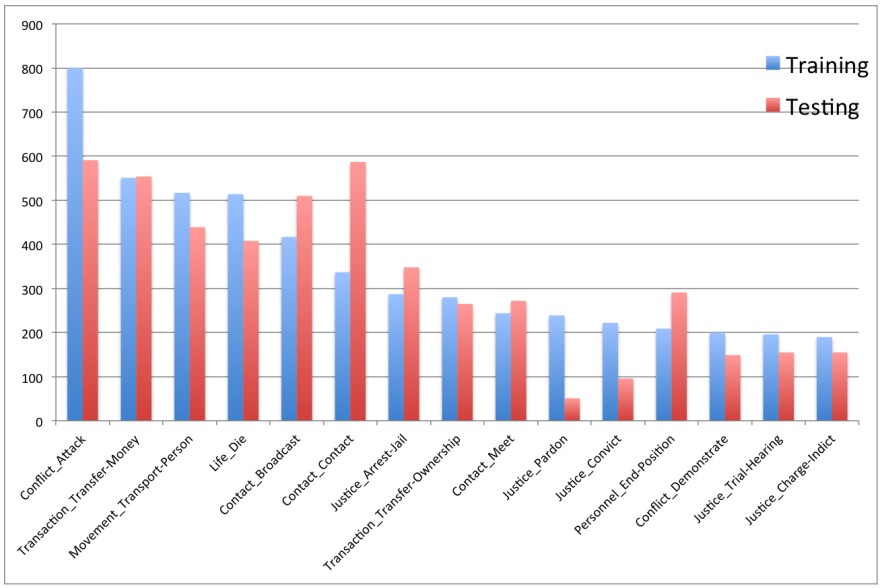

Figure 1: Number of mentions by type: Training vs. Test

10 Appendix

10.1 Team Name Mapping

The Team ID used in this paper refers to the team listed in Table 12.

Team1	BUPT_PRIS
Team2	CMU_CS_event
Team3	HITS
Team4	IHMC
Team5	LCC
Team6	LTI
Team7	NTNU
Team8	OSU
Team9	RPI_BLENDER
Team10	SYDNEY
Team11	TEA_ICT
Team12	UI_CCG
Team13	UKP
Team14	UMBC
Team15	UTD
Team16	WIP
Team17	ZJU_Insight

Table 12: Team ID list

References

A. Bagga and B. Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 79–85. Association for Computational Linguistics.

Nancy Chinchor. 1992. MUC-5 EVALUATION METRIC. In *Proceedings of the 5th Conference on Message Understanding*, pages 69–78.

Language Technologies Institute - Carnegie Mellon University. 2015a. Event Nugget Detection and Coreference Scoring v.27. Technical report.

Language Technologies Institute - Carnegie Mellon University. 2015b. TAC KBP Event Detection Annotation Guidelines v1.7. Technical report.

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.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, (October):25–32.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*, pages 55–60.

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.

M Recasens and E Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 1(1).