Raytheon
BBN Technologies

TAC KBP:
Event Argument Extraction & Linking

Marjorie Freedman, Ryan Gabbard,
Yee Seng Chen, Jay DeYoung

The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Distribution Statement "A" (Approved for Public Release. Distribution Unlimited.)

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) DEFT Program.

TASK OVERVIEW

Event Nugget Detection & Coreference (EN) Task

Raytheon BBN Technologies

- Detect, label and coreference spans of text that indicate the presence of an event (the event nugget) in the ontology
  - *"A separatist group called TAK claimed responsibility for an explosion which wounded six people… Istanbul governor Muammer Guler told Anatolia news agency the explosion injured six people. "*
    - *explosion → Conflict.Attack(corefID = 1)*
    - *wounded → Life.Injure (corefID =2),*
    - *explosion → Conflict.Attack (corefID=1)*
    - *injured →Life.Injure (corefID=2)*
  - Also reported realis status for the event (ACTUAL, GENERIC, OTHER)

Event Argument Extraction Linking (EAL) Task

Raytheon BBN Technologies

- For a set of documents
  - Identify what events occurred along with their type
  - Identify key arguments (e.g. participants, dates, locations) and associate them with the correct events
  - Provide realis status (ACTUAL, OTHER, GENERIC)

A separatist group called the Kurdistan Freedom Falcons (TAK) claimed responsibility for an explosion late on Monday which wounded six people, one of them seriously, in an Istanbul supermarket. Istanbul governor Muammer Guler told Anatolia news agency the explosion in the Bahcelievler district of Turkey's largest city injured six people. The agency said 15 other people had been hurt. "We consider the explosion that took place tonight in an Istanbul supermarket to be a response to the barbaric policies against the Kurdish people

Event2: Conflict.Attack	Role	Fillers
	ATTACKER	TAK
	TARGET	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

Event1: Life.Injure	Role	Fillers
	Agent	TAK
	Victims	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

Raytheon
BBN Technologies

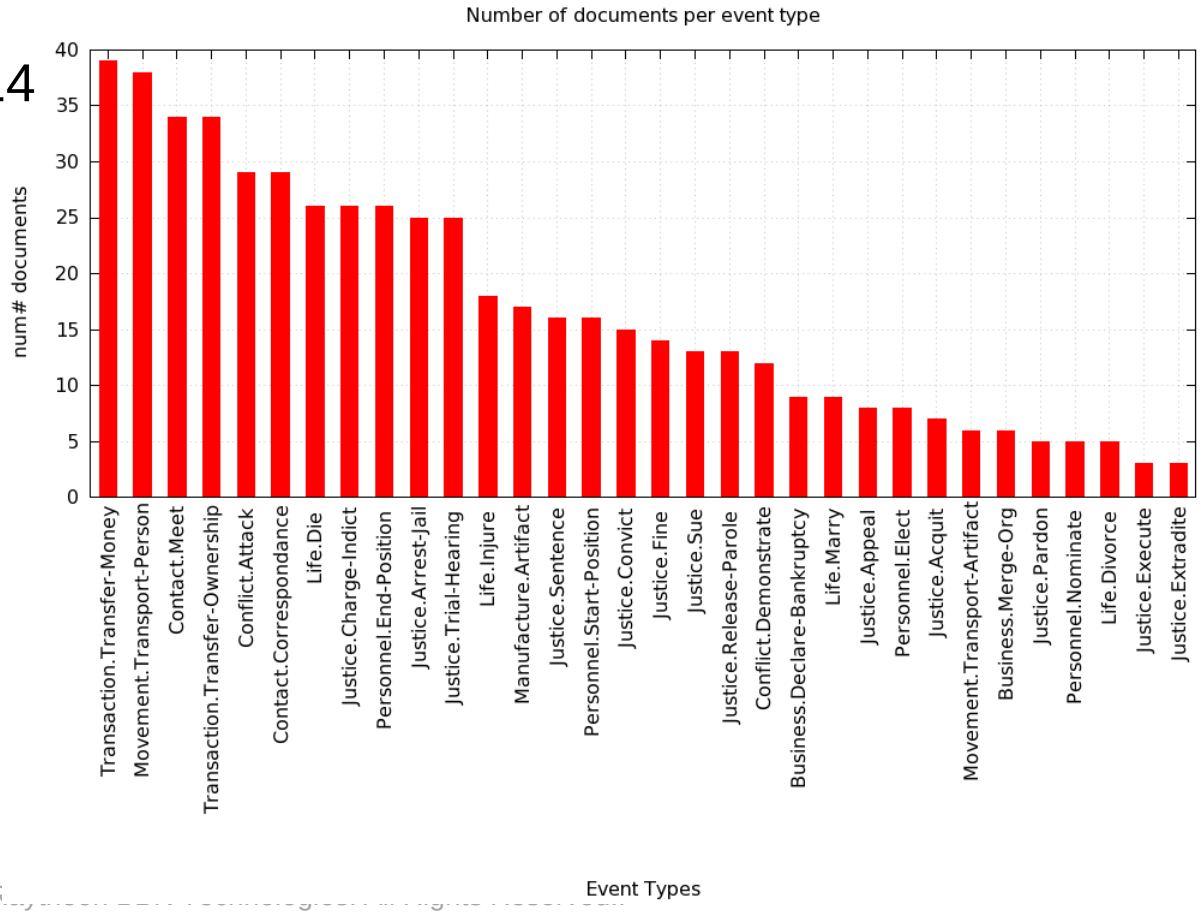
What is Required to Fill an Event Frame

1. Finding events, arguments, and their roles (2014 task)
   A. Recognize the presence of the event → *overlap with the event nugget task but no requirement that the exact phrase is found; instead allow sentence length justifications*
   B. Find a mention (base filler) where the participation in the event (along with the role) is clear → *similar to mention level argument extraction as in event detection in ACE*
   C. Link the base filler to a canonical argument string → *use within document coreference and temporal resolution; similar to ColdStart requirement that slot-fills reference a named entity (and not a local mention)*
   D. Assign a realis label to assertion about the event and argument → *overlap with the event nugget task, but also incorporate understanding of the argument itself (e.g. failed participation)*
2. Link the argument assertions such that arguments that correspond to the same "real world" event are grouped together (new in 2015)

Event Ontology

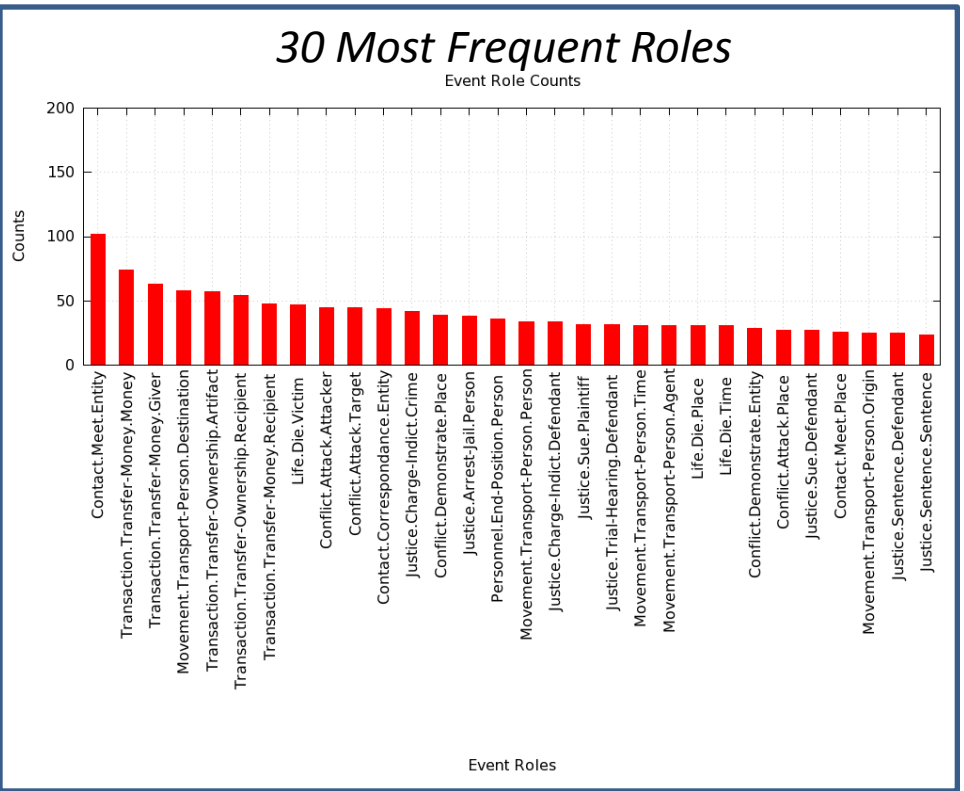
- Rich ERE event ontology (similar to ACE, TAC 2014)
  - EAL: As in 2014, ignore events for which all arguments are subsumed by the ColdStart/SlotFilling evaluation
    - Life.Be-Born
    - Business.Found
- Changes between EA 2014 and TAC 2015
  - RichERE subtypes for
    - Contact (EAL: Ignore Contact.Broadcast)
    - Movement
    - Transaction
  - From RichERE, add Manufacture.Artifact

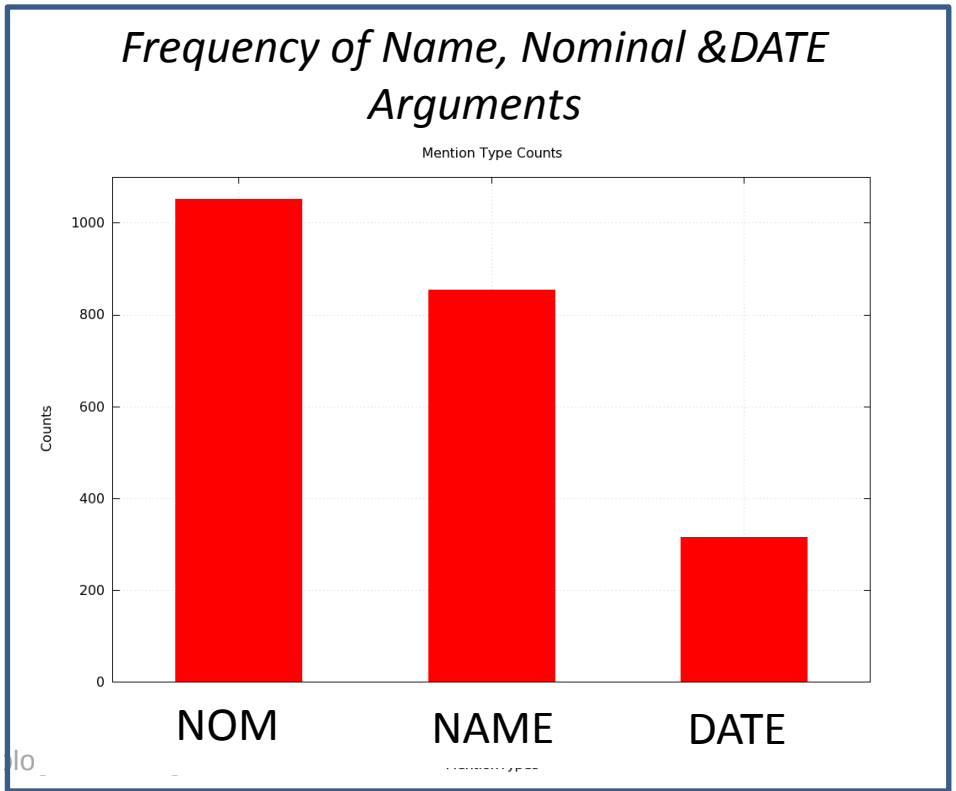
EAL: Event Frequency in 81 Assessed Documents



Event Ontology: Arguments (EAL Only)

- Each event class has
  - A set of ontology driven argument roles (recipient, artifact, crime etc.)
  - General date/location arguments
- Arguments can be named or non-named (e.g. the crowd, the unnamed suspect)
  - Arguments include entities (people, facilities), dates, and non-entity fills (crime, sentence)





Raytheon
BBN Technologies

TAC KBP:
Event Argument Extraction & Linking

Marjorie Freedman, Ryan Gabbard,
Yee Seng Chen, Jay DeYoung

Raytheon
BBN Technologies

LDC DATA OVERVIEW

Event Argument Extraction Linking (EAL) Task

Raytheon BBN Technologies

- For a set of documents
  - Identify what events occurred along with their type
  - Identify key arguments (e.g. participants, dates, locations) and associate them with the correct events
  - Provide realis status (ACTUAL, OTHER, GENERIC)

A separatist group called the Kurdistan Freedom Falcons (TAK) claimed responsibility for an explosion late on Monday which wounded six people, one of them seriously, in an Istanbul supermarket. Istanbul governor Muammer Guler told Anatolia news agency the explosion in the Bahcelievler district of Turkey's largest city injured six people. The agency said 15 other people had been hurt. "We consider the explosion that took place tonight in an Istanbul supermarket to be a response to the barbaric policies against the Kurdish people

Event2: ATTACK	Role	Fillers
	ATTACKER	TAK
	TARGET	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

Event2: ATTACK	Role	Fillers
	Agent	TAK
	Victims	Six people 15 other people
	PLACE	the Bahcelievler district Istanbul An Istanbul supermarket
	DATE	Monday (2006-02-13)

Raytheon
BBN Technologies

Tasks & Participants

- 2015: Two subtasks
  - **EAL**: Given new documents, produce event frames for each document
    - LDC produces a manual run for this task, with annotators spending 60 minutes per document
    - system participants
      - 1 participant submitted results with incorrect offsets→
        - » Results not included in preliminary results, but will be integrated later
        - » On smaller sample, missing system is below median
  - **EVL**: Given all system argument responses, perform system combination/filtering and linking
    - 1 participant, results are reported with EAL results

Task Data

- Participants were encouraged to use a mix of existing resources as training
  - ACE event annotation
  - Rich ERE event annotation
  - Event Nugget training
  - Assessments from 2014
- LDC provided a small amount of task specific training data: linking arguments from 50 files of 2014 assessments
- Participants produced output for 500 English documents, evenly split between discussion forum (DF) and newswire (NW)
  - 81 of these documents have been assessed
  - Many documents overlap with the Event Nugget documents

Scoring (1)

- Overall metric combines accuracy of the quality of event argument extraction (EaArg) with accuracy of linking these arguments (EaLink)
  - EaArg: $TruePositive_{EAE} - \beta FalsePositive_{EAE}$, β set to 0.25
    - Report F1 for arguments as additional diagnostic
  - EaLink: B^3
    - Modified to: (1) Ignore system false alarm arguments; (2) Allow an argument to appear in multiple clusters
    - EaLink is sensitive to system recall-- a system with low argument recall will have a low maximum EaLink score
- *TruePositives* and *FalsePositives* are calculated on entity/real world objects (and not for each mention of an object)
  - Example: *"Sue and Bob attended a meeting. Sue gave a presentation at the meeting. She brought handouts for the meeting."*
    - EAL Scoring
      - One TruePositive for Sue (as a meeting attendee)
      - One TruePositive for Bob (as a meeting attendee)
    - *ACE event mention argument scoring*
      - Three TruePositives for Sue (as a meeting attendee)
      - One TruePositive for Bob (as a meeting attendee)

Description of Scoring: http://www.nist.gov/tac/2015/KBP/Event/Argument/guidelines/EventArgumentExtractionandLinkingScoringProposal.v07.pdf
Scorer code: http://github.com/rgabbard-bbn/kbp-2014-event-arguments

Scoring (2)

- In official score, for a response to be correct all aspects of the argument assertion must be assessed as CORRECT or INEXACT,
  - This includes
    - Justification that the event is present
    - Justification that the entity involved participated in the event
    - Resolution to a canonical string
    - Assignment of a realis label
  - Requirement that all aspects of an assertion be CORRECT/INEXACT reflects KB accuracy of event-argument related assertions
- Diagnostic analysis relaxes these constraints
    - Realis label is ignored
    - Impact of resolution to a canonical name is ignored
  - More relaxed measure is more tightly correlated with core improvements to finding and labeling event arguments

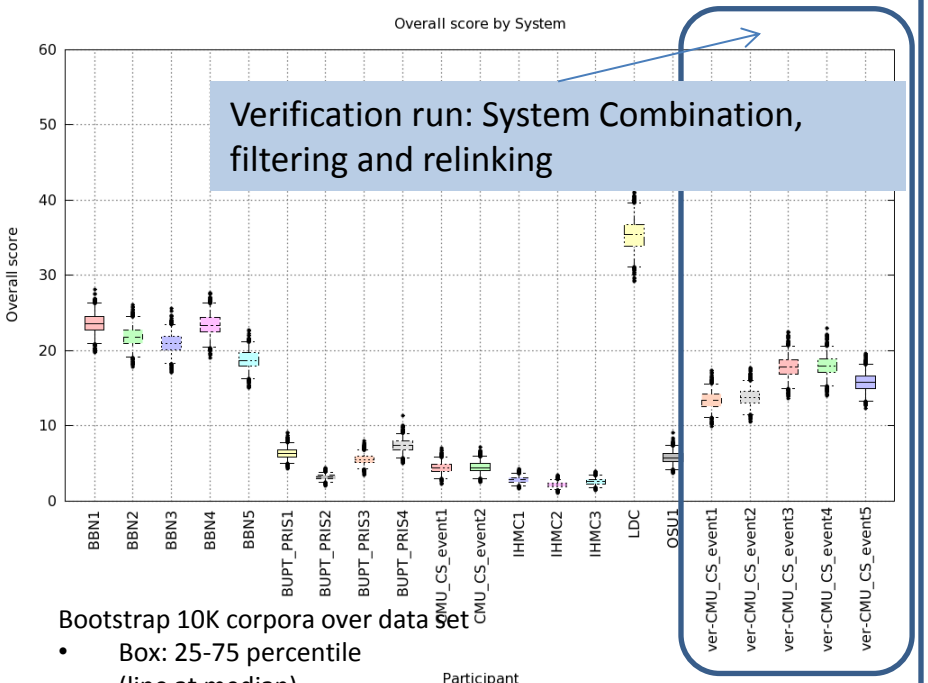
Preliminary Results: Overall Metric

- Results still preliminary
  - Will add results from an additional system and handful of additional files
  - Some additional QC
- LDC performance exceeds all systems
- Large gap between rank 1 system and other systems
  - One team submitted a “verification and linking run”– Overall performance does not exceed Rank1 system

Metrics: Top Submission Per-Team

Submission	P	R	F1	EAArg	EALink	Overall
LDC	76	40	52	37	34	35
BBN1	***37***	***39***	***38***	***24***	***23***	***24***
ver-CMU_CS_event4	32	38	35	20	17	18
BUPT_PRIS4	30	16	21	8	7	7
OSU1	24	15	18	6	6	6
CMU_CS_event2	31	10	15	5	4	4
IHMC1	10	13	11	1	4	3

Overall Score: All Submissions



Bootstrap 10K corpora over data set

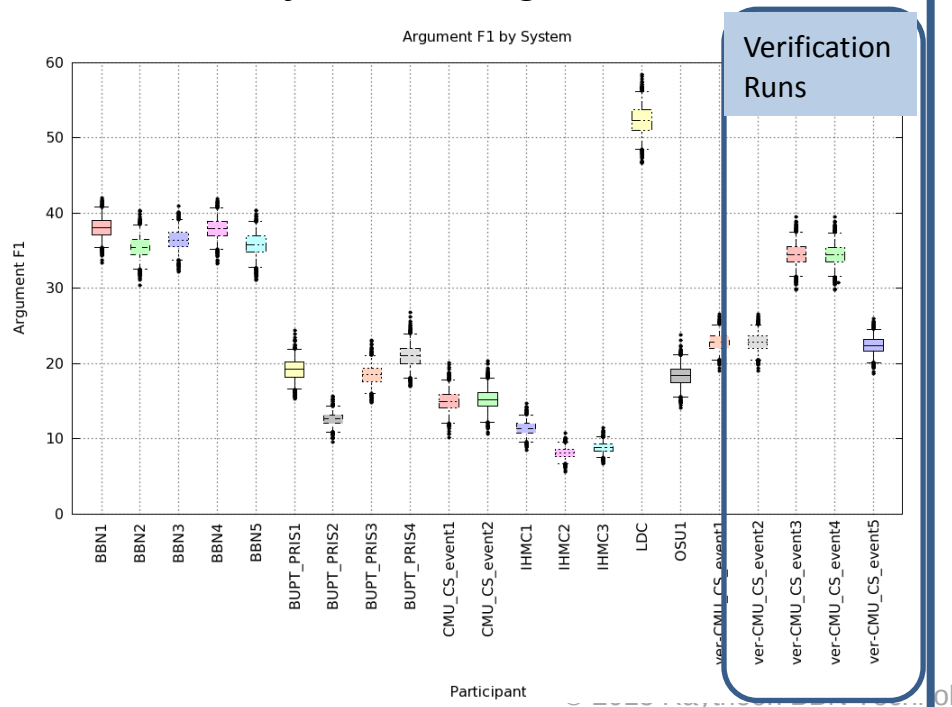
- Box: 25-75 percentile (line at median)
- Whisker: 5-95 percentile
- Points: outliers

Argument Score (1)

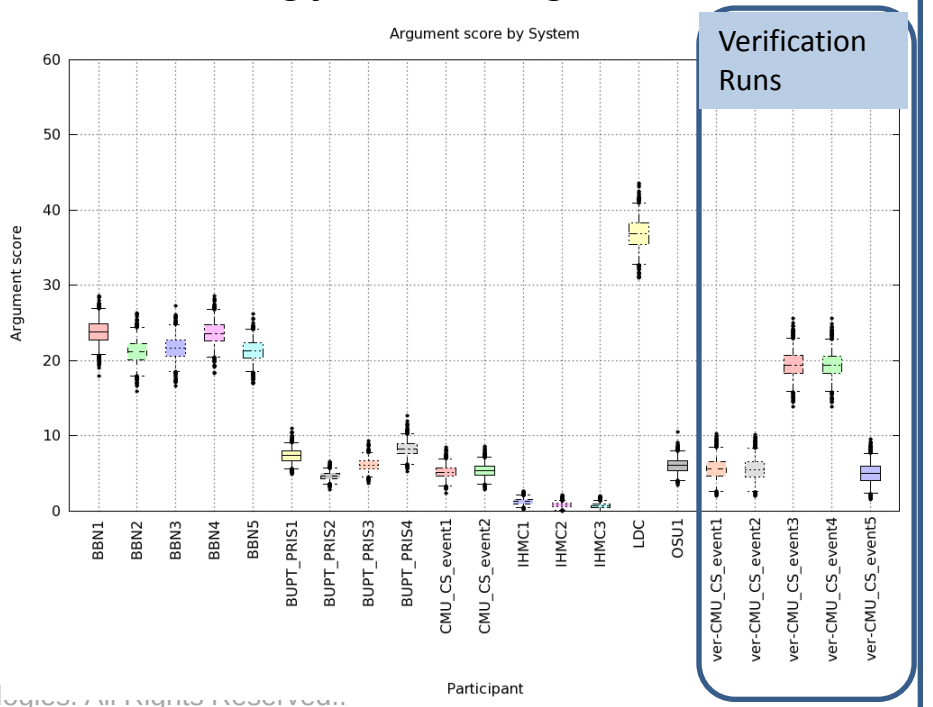
- Absolute F1 is higher than EAArg score
  - Rankings fairly stable

Submission	F1	EAArg
LDC	52	37
BBN1	***38***	***24***
ver-CMU_CS_event4	35	20
BUPT_PRIS4	21	8
OSU1	18	6
CMU_CS_event2	15	5
IHMC1	11	1

F1 for Event Arguments



EAArg for Event Arguments



Compare 2014 EA with 2015 EA

Raytheon BBN Technologies

Sub.	2014: Top Submission Prec	2015: Prec	Rel. Change: Prec	2014: Top Submission Rec	2015: Rec	Rel. Change: Rec	2014: Top Submission F1	2015: F1	Rel. Change: F1
LDC	75.88	76.27	1%	27.93	34.81	25%	40.83	47.8	17%
BBN1	44.66	36.83	-18%	25.9	39.49	52%	32.79	38.11	16%
BBN4		36.95	-17%		39.02	51%		37.96	16%
BBN2		34.22	-23%		36.95	43%		35.53	8%
BBN3		36.97	-17%		36	39%		36.48	11%
BBN5		46.14	3%		29.33	13%		35.86	9%
OSU1	29.11	24.05	-17%	22.58	14.94	-34%	25.43	18.43	-28%
IHMC1	11.41	10.08	-12%	6.1	13.01	113%	7.95	11.36	43%
IHMC3		6.55	-43%		13.65	124%		8.85	11%
IHMC2		6.82	-40%		9.99	64%		8.11	2%

- Compare scores across years for teams that participated in both years
  - For comparison, use a site's top 2014 submission
  - 2015 argument task added new event types, so an unmodified 2014 system would have performed less well on the 2015 task
  - Different data set: Distribution of event types is different in 2014 than 2015
- Two sites improve over 2014 performance (BBN, IHMC)
  - For the top performing instance of both, recall improves but precision decreases
  - BBN5 (a lower ranked BBN submission) improves precision and recall

Raytheon
BBN Technologies

ANALYSIS OF ARGUMENT EXTRACTION: RESULTS & APPROACHES

Data Resources

- Participants self-report the training/knowledge resources they use. Table: Per-site resources used in order of ranked overall score
  - Most sites used the same resources for all runs
  - BBN2 differed from other BBN runs (by excluding internally created training data)
- Top ranked system made most extensive use of resources
- Background corpus was used by
  - BBN → Brown clusters (all runs), Contextual Embeddings (BBN1, BBN2, BBN4, BB5)
  - CMU_CS_Event → Embeddings

Run ID	ACE2005	RichERE	Event Nugget Training	EA Assmnt	EAL Training	WordNet	Background Corpus	Internal Training Data	Internal Assmnts	Other
BBN1,3,4,5	Yes	Yes	No	Yes	Yes	Yes	Yes (Gigaword)	Yes	Yes	
BBN2	Yes	Yes	No	Yes	Yes	Yes	Yes (Gigaword)	No	Yes	
BUPT_PRIS1-4	Yes	No	No	Yes	Yes	No	No	No	No	
OSU1	Yes	Yes	No	No	No	Yes	No	No	No	
CMU_CS_event1-2	Yes	No	No	No	No	No	Yes (English Wikipedia)	No	No	
IHMC1-3	No	Yes	No	Yes	Yes	No	Yes (English Wikipedia)	No	No	VerbNet, CatVar
ZJU_Insight1-5	Yes	No	Yes	No	Yes	Yes	No	No	No	
ver-CMU_CS_event1-5	Yes	No	No	No	No	No	No	No	No	

Software Resources

- Participants self-report use of NLP tools. Table: Per-site tools used in order of ranked overall score
- All systems used NER , top ranked system was only site to report using a nominal classifier (*nominal classification: city → GPE; 7 victims → PER)*
  - BBN used entity types of nominals to constrain argument decisions
  - Other systems may have used this information via CoreNLP coreference chain
- Several systems reported ignoring temporal resolution– will result in incorrect CAS assessments for DATE arguments
- All systems used some form of entity coreference
  - Necessary for finding the canonical argument string for an argument
    - "They held signs protesting the initiative" → *(Conflict.Demonstrate, Agent, "group of students")*
  - Most systems used CoreNLP
- All systems used some form of syntactic parsing
  - Only one system reported using semantic role labeling

Run ID	NER	Nominal Classification	Temporal Resolution	InDoc Coref	Syntactic Parser	Other
BBN	Yes	Yes	Yes	Yes (BBN SERIF)	Yes (BBN SERIF)	No
BUPT_PRIS	Yes	No	No	Yes (CoreNLP)	Yes (CoreNLP)	Yes (CoreNLP?)
IHMC	Yes	No	No	Yes (CoreNLP)	Yes (CoreNLP)	Yes (Semantic Role Labeler)
OSU	Yes	No	Yes	Yes (CoreNLP)	Yes (CoreNLP)	Yes (CoreNLP?)
CMU_CS_event	Yes	No	Yes	Yes (JET)	Yes	Yes (JET, CoreNLP?)
ZJU_Insight	Yes	No	No	Yes (CoreNLP)	Yes (CoreNLP)	No

Approaches to Finding Events & Arguments

- Most systems used a trained classifier to find and label the type of event
  - At least one system (IHMC) also explored manually crafted rules
  - Some systems reported classifying at the sentence level, others at the nugget/trigger level
- Most systems used a trained classifier to label arguments
  - Features include the entity types of arguments, syntactic context, lexical context, etc.
  - BBN used heuristics to propagate arguments from one event to another (specifically for violent events)

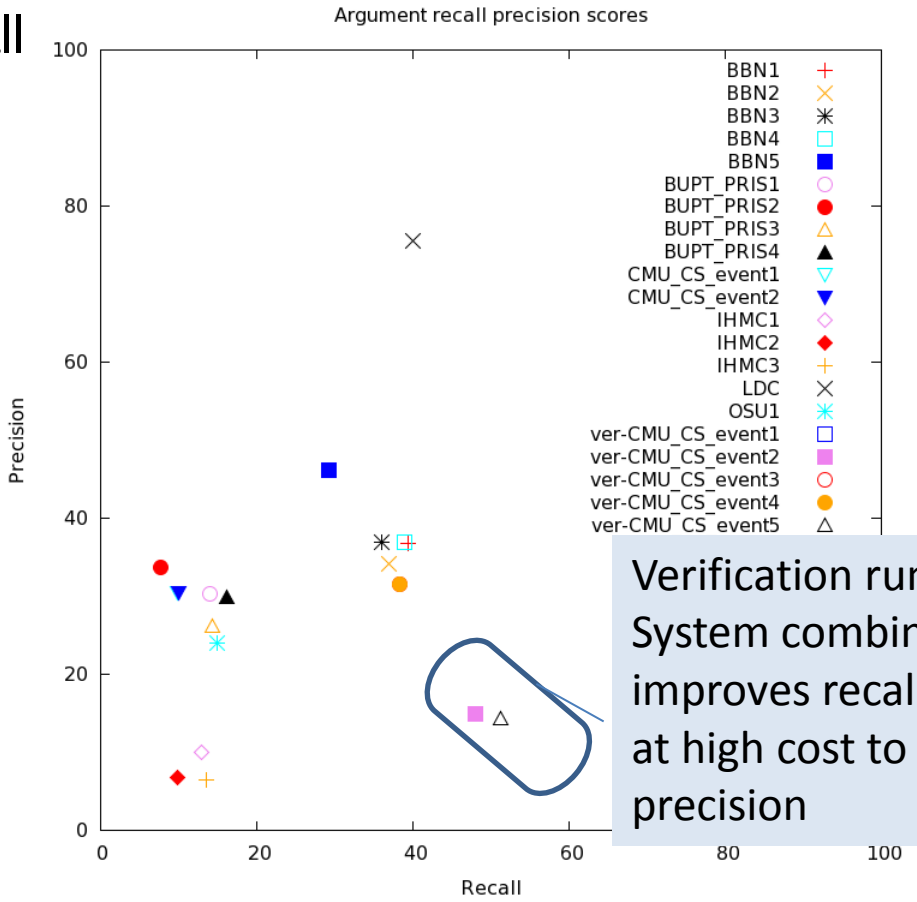
Approaches to Labeling Realis & Providing Confidence

Raytheon BBN Technologies

- Realis (distinguishing ACTUAL, GENERIC, OTHER)
  - 3 sites used trained classifiers for labeling realis
    - Features include tense of verb(s), features of the argument
    - 3 sites used ACE for training, 1 used 2014 EA assessments
  - 1 site (BBN) used a mix of rules and a trained classifier
  - 1 site (IHMC) used purely linguistic rules
- Only two sites self-reported producing meaningful confidences (BBN, CMU)

Arguments: Precision and Recall

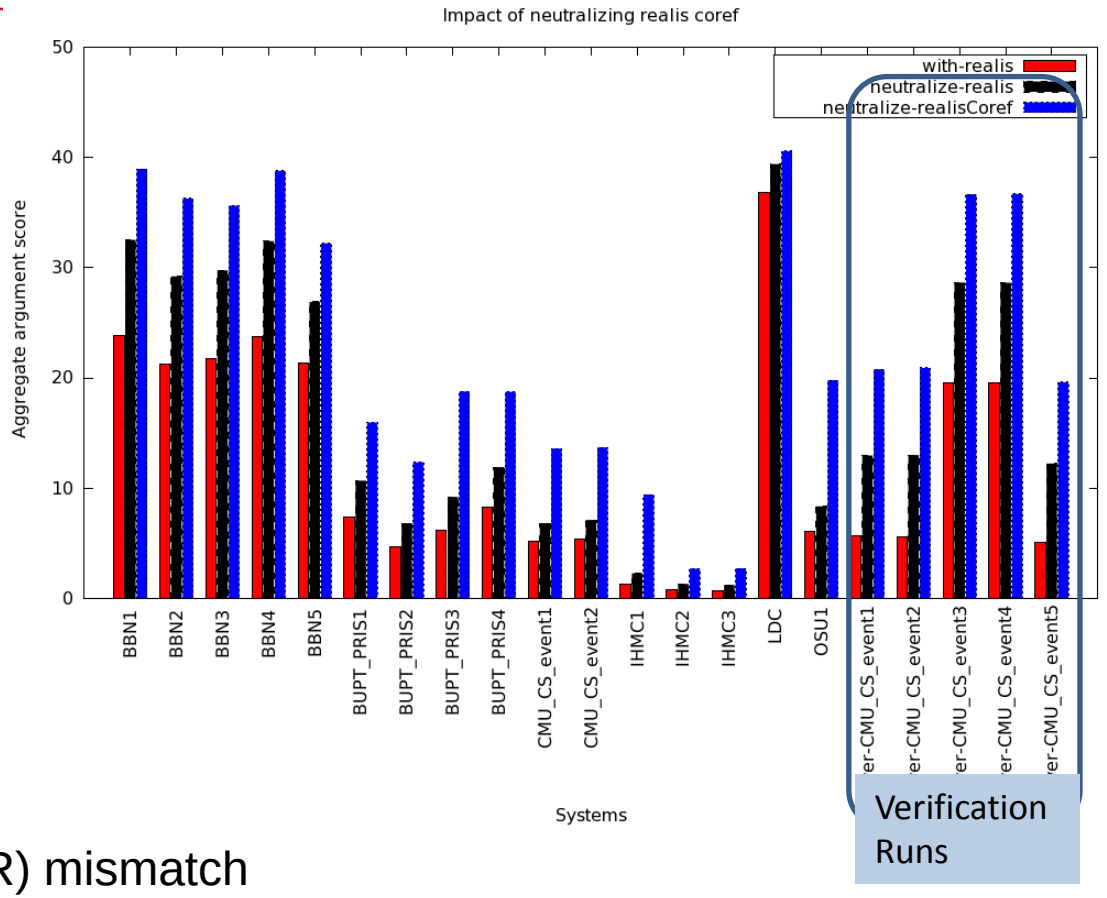
- Most systems operating at points of higher precision than recall
  - Top system (BBN1) balances precision &recall



Verification runs -- System combination improves recall, but at high cost to precision

Diagnostic Measure

- **Official metric (with-realis[red]):** treats incorrect realis & canonical argument resolution as an error
  - Realis/CAS distinct subsystems for all participants
- Diagnostic scoring: Focus on finding and labeling EA tuples without needing to make complete KB assertions
  - **neutralize-realis: (black)** Treat assertion as correct regardless of (ACTUAL, GENERIC, OTHER) mismatch
  - **neutralize-realisCoref (blue):** Neutralize realis and use heuristics to align system mention (base filler) with a correct canonical argument string
- With diagnostic scoring, all systems improve (a lot)



Diagnostic Measure(2)

- All submissions see some improvement from diagnostic measurement
  - For top system (BBN1) relative impact of realis seems larger than that of coreference
    - BBN1 uses internally developed coreference and incorporates coreference confidence into decision about whether or not to report an argument
  - For other EA systems, relative impact of coreference is greater than that of realis
    - Systems that use CoreNLP coreference as a black box may be hindered in improving EA-KB assertions
- Within document coreference is not formally evaluated in TAC KB, but without improvements to within document coreference tasks suffer

	NeutralizeRealis improvement over *Official*			*NeutralizeRealisCoref* improvement over *NeutralizeRealis*			*NeutralizeRealisCoref* improvement over *Official*		
	Rel Imp P	***Rel Imp R***	***Rel Imp EAArg***	***Rel Imp P***	***Rel Imp R***	***Rel Imp EAArg***	***Rel Imp P***	***Rel Imp R***	***Rel Imp EAArg***
LDC	8%	4%	7%	3%	2%	3%	12%	6%	10%
BBN1	23%	16%	37%	15%	10%	20%	42%	27%	64%
ver-CMU_CS_event4	23%	20%	47%	18%	12%	28%	45%	34%	88%
BUPT_PRIS4	21%	25%	44%	32%	26%	57%	60%	57%	126%
OSU1	20%	24%	37%	57%	52%	137%	89%	88%	224%
CMU_CS_event2	16%	19%	32%	48%	44%	92%	72%	71%	154%
IHMC1	54%	8%	78%	62%	55%	312%	150%	68%	631%

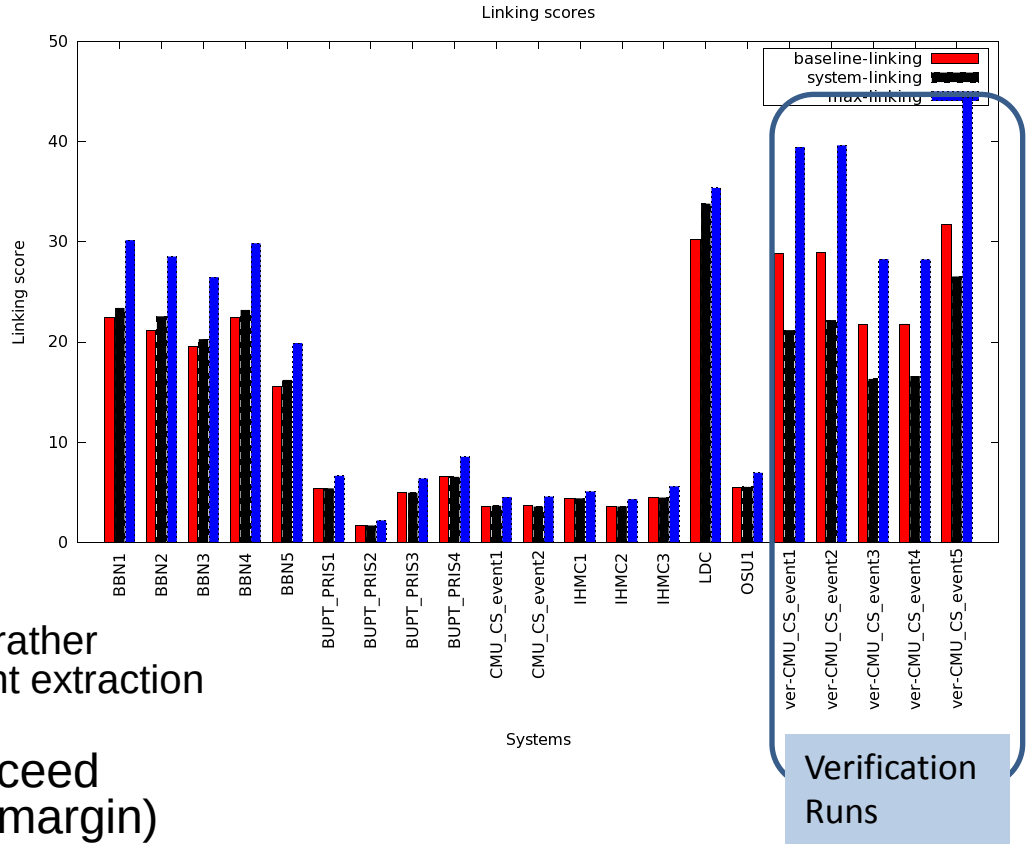
EVENT ARGUMENT LINKING

Event Argument Linking

- Linking arguments into event frames new in 2015
  - Event argument frames were defined to be at the same level of granularity as ERE Event Hoppers
  - Implemented baseline of “link all arguments in a document with the same event type” was available to all participants
- Maximum linking score (EALink) is bounded by a submission’s argument recall
  - Systems that don’t find arguments can’t produce complete event argument frames
- Participant approaches to linking
  - BBN used a sieve based approach
  - Two systems used the baseline approach (BUPT, IHMC)
  - CMU approach linked arguments based on the their trigger
    - This may have limited linking to sentence internal links
    - This approach was also used in the verification and linking task
  - OSU integrates left linking event coreference for triggers

Linking: Comparing to Baseline

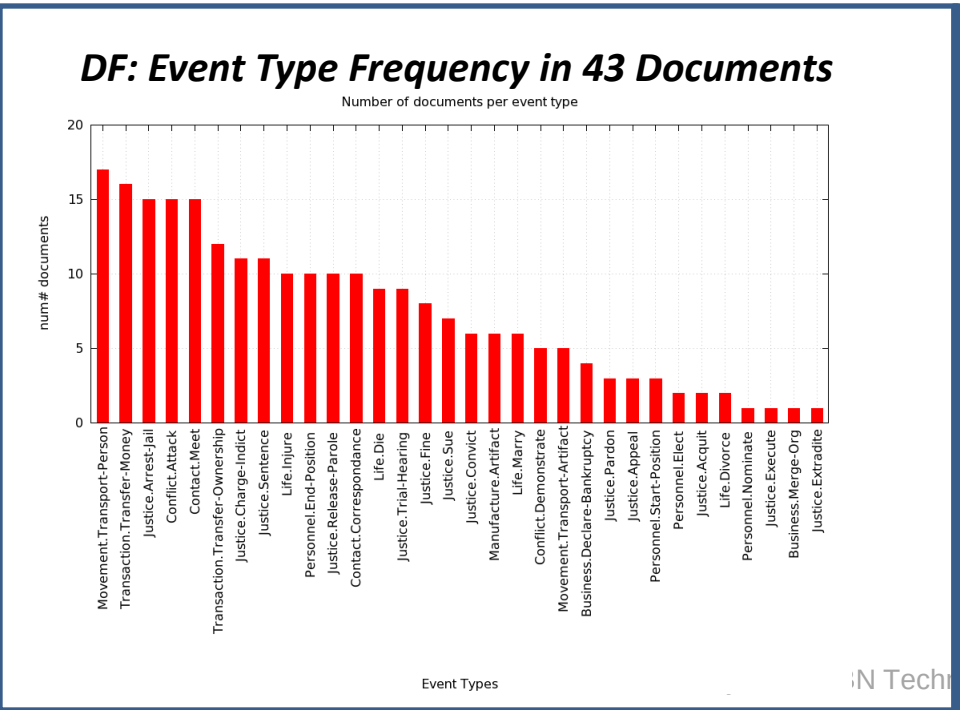
- Compare
  - **System-linking (black):** Accuracy submitted linking
  - **Baseline-linking (red):** Accuracy of linking with baseline
    - Used by BUPT & IHMC
  - **Max-linking (blue):** Given arguments found by a system, what is the best linking score it could get
- For all but the top site (BBN), difference between baseline and max linking is small
  - Reflects low recall of other systems
  - To make progress linking over system (rather than perfect) arguments, basic argument extraction needs to improve
- Only BBN submissions noticeably exceed baseline linking (and only by a small margin)
  - BBN observed a much smaller difference between baseline and max linking in the 50 EAL-specific training documents
- CMU verification run used a linking strategy that was likely to link only sentence internal arguments
  - On CMU submission (which is low recall), this strategy is roughly equivalent to the baseline
  - On verification submissions, (where many more arguments were available) this approach falls well below baseline

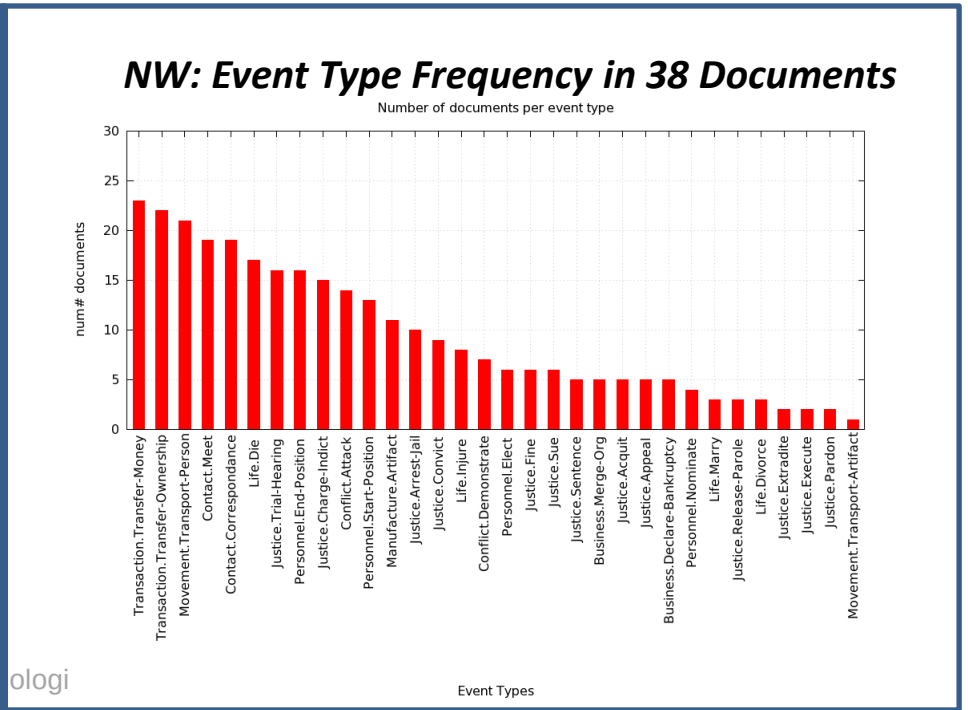


GENRE DIFFERENCES

Task Data: Differences in DF/NW

- Roughly equivalent number of documents in DF and NW set, but far few correct argument assertions in DF
  - DF: 883
  - NW: 1,439
- Most event types that are high frequency in one genre are high frequency in the other
    - A few exceptions

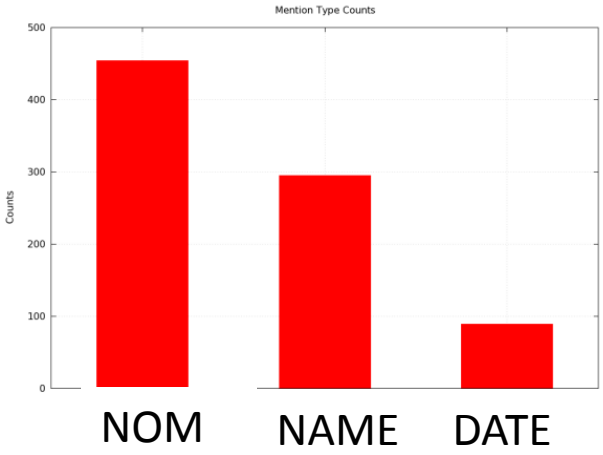




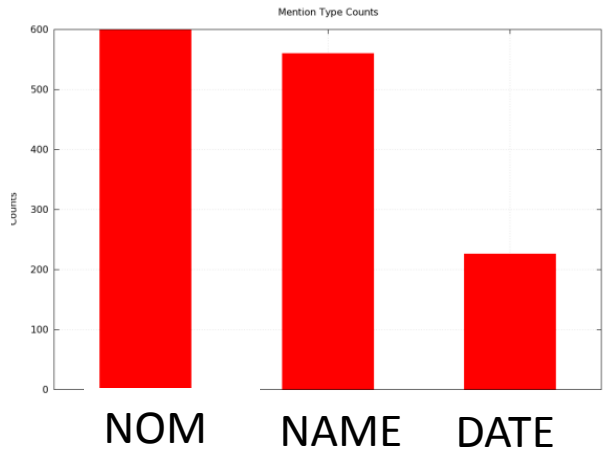
Data: NW and DF

- Both genre: nominals most common (Especially DF)

Discussion Forum

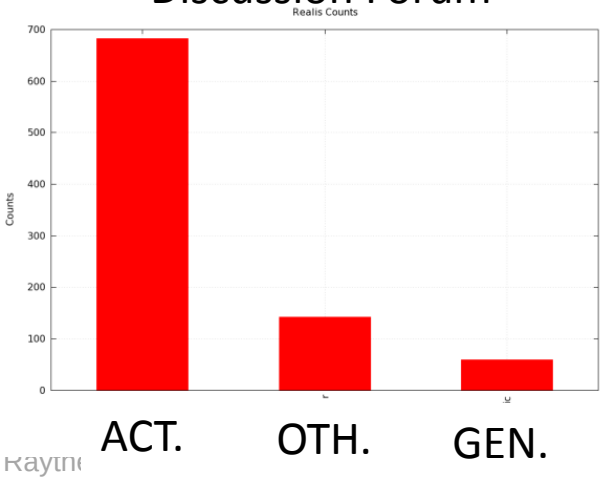


Newswire

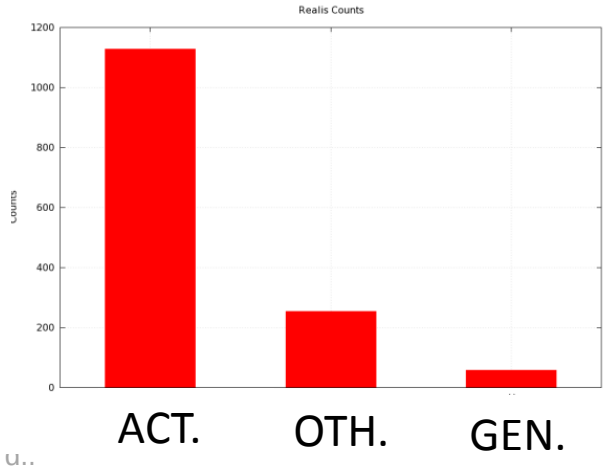


- Both genre: ACTUAL is far more common than OTHER or GENERIC

Discussion Forum



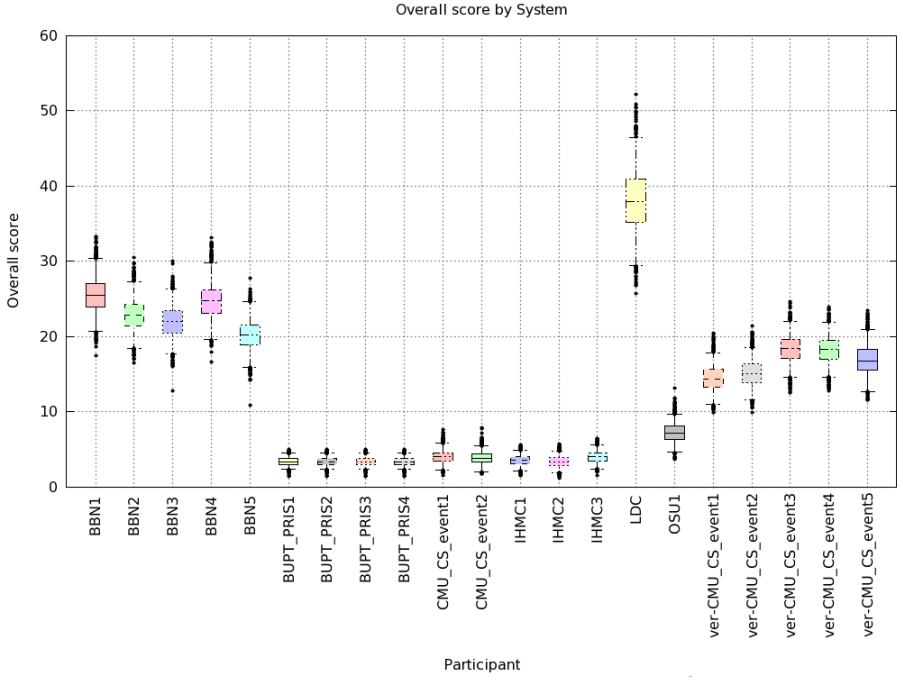
Newswire



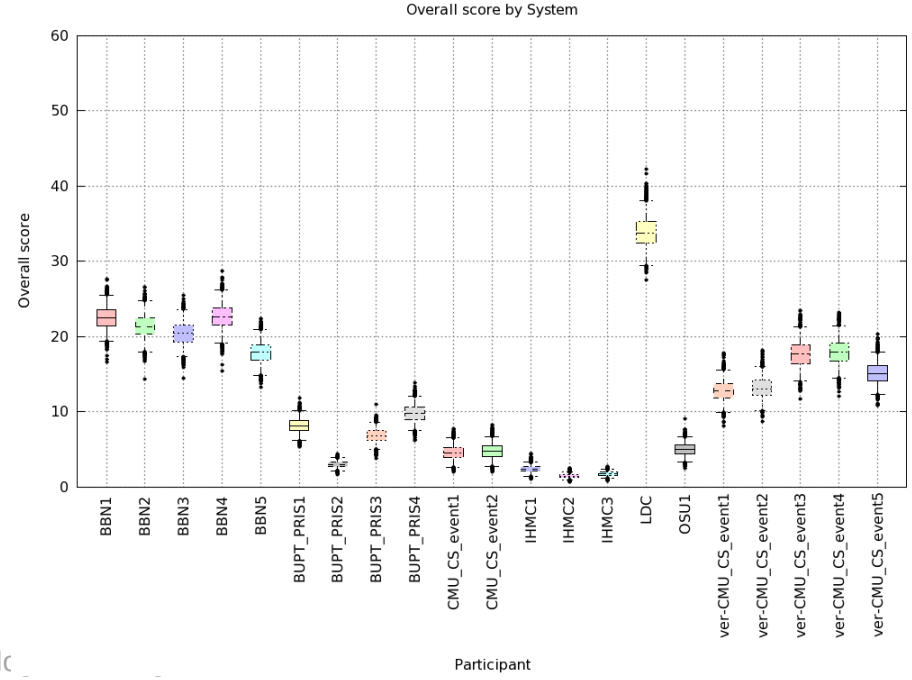
Genre Differences: Performance

- LDC and Rank 1 system (BBN1) performed better on DF than NW
- Bootstrap sampling shows more variation of the expected-performance for LDC and BBN1 on DF than NW

Overall Score: Discussion Forum Data



Overall Score: Newswire Data



Raytheon
BBN Technologies

Preview of 2016 Discussion

- Multilingual (English, Chinese, Spanish)
- Cross-document

- https://www.surveymonkey.com/r/RT99HS9

- Thanks!