Raytheon
BBN Technologies

BBN System

TAC KBP 2015, Event Argument Linking

Marjorie Freedman

Ryan Gabbard

Jay DeYoung

Yee Seng Chan

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.

Outline

- Overview of our system
- New for our 2015 system
  - Targeted training (trigger)
  - Embeddings (argument attachment)
  - Sieve (argument linking)
- Submissions, results, and analysis

Task Output

Raytheon
BBN Technologies

- Labels from taxonomy: Event Type (ET) and Role (AR)
- Canonical Argument String (CAS): Best resolution for the argument
- Elements of the justification
  - **Base Filler (BF):** Span of text that fills the specified role
  - **Predicate Justification (PJ):** Span of text that indicates the presence of an event of the specified type and the participation of the base filler
  - **Additional Argument Justifications:** Span of text that establishes the relationship between the base filler and canonical argument string for non-coreference relations (e.g. group membership)
- Realis Marker: Categorization of epistemic status of (EventType, Role, Cannonical Argument) assertion. Labels are {Actual, Other, Generic}
- Confidence: System confidence in (EventType, Role, Cannonical Argument) assertion

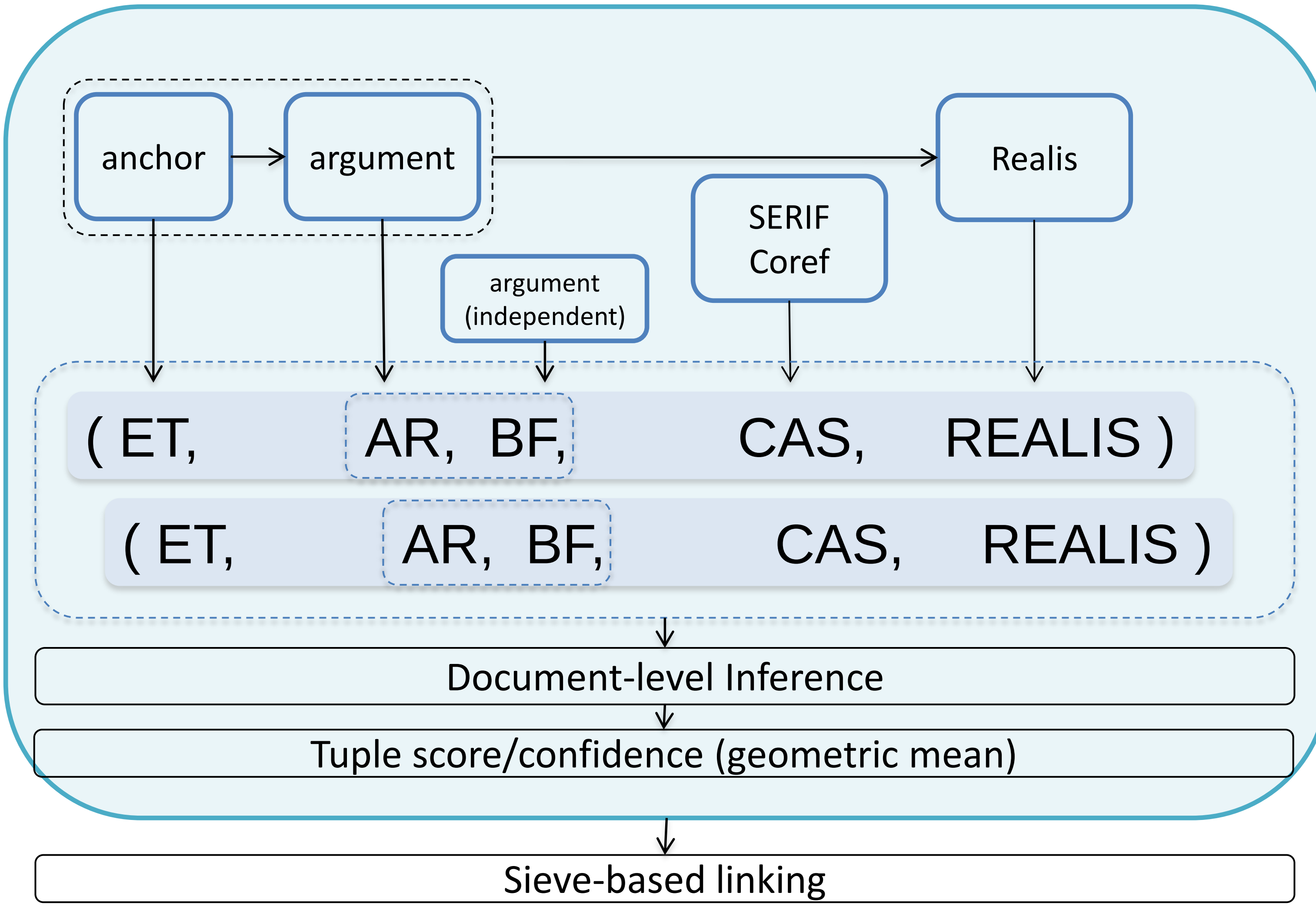
Prosecutors in the Oscar Pistorius case said they will file appeals. The Olympic runner was convicted of killing his girlfriend.

	ET	Role	CAS	Realis	Justifications (base filler is underlined)
Ev1	Appeal	Prosecutor	Prosecutors	Other	*Prosecutors in the Oscar Pistorius case said they will file appeals.*
	Appeal	Defendant	Oscar Pistorius	Other	*Prosecutors in the Oscar Pistorius case said they will file appeals.*
	Appeal	Crime	killing his girlfriend	Other	Prosecutors in the Oscar Pistorius case said they will file appeals. The Olympic runner was convicted of killing his girlfriend.
Ev2	Convict	Defendant	Oscar Pistorius	Actual	The Olympic runner was convicted of killing his girlfriend.
	Convict	Crime	killing his girlfriend	Actual	The Olympic runner was convicted of killing his girlfriend.

BBN System

Core Components Similar to EA 2014

New for 2015



Anchor Model

- Finds lexical indications of events (anchors/triggers/nuggets), and label with event type
- Logistic regression
  - Training data: ACE, RichERE
    - In BBN1, 3,4,5 also includes BBN developed targeted training data (new in 2015)
  - Features:
    - Surrounding words, associated dependency structure, topic of document, etc.

Argument & Argument Independent Models

- **Argument Model:**
  - Given the presence of a lexical indicator of the event (from the anchor model) and some noun phrase, predict the role (if any)
  - Logistic regression (more complicated factor for embeddings)
    - Training data: ACE, RichERE
    - Features include:
      - The anchor, the candidate argument, the text strings between them, and their associated dependency structures
      - In BBN1,2,4,5: embedding features (new in 2015)
- **Argument Independent Model:**
  - Given a mention, predict presence of an event type and role that mention plays in event
    - Training data: ACE, RichERE
    - Features:
      - As in the core argument model, but without the anchor
  - Targets improved recall for mentions that are highly indicative of certain classes of events (e.g. *the protestors* → *(Conflict.Demonstrate, Agent, the protestors)*

Canonical Argument String (CAS)

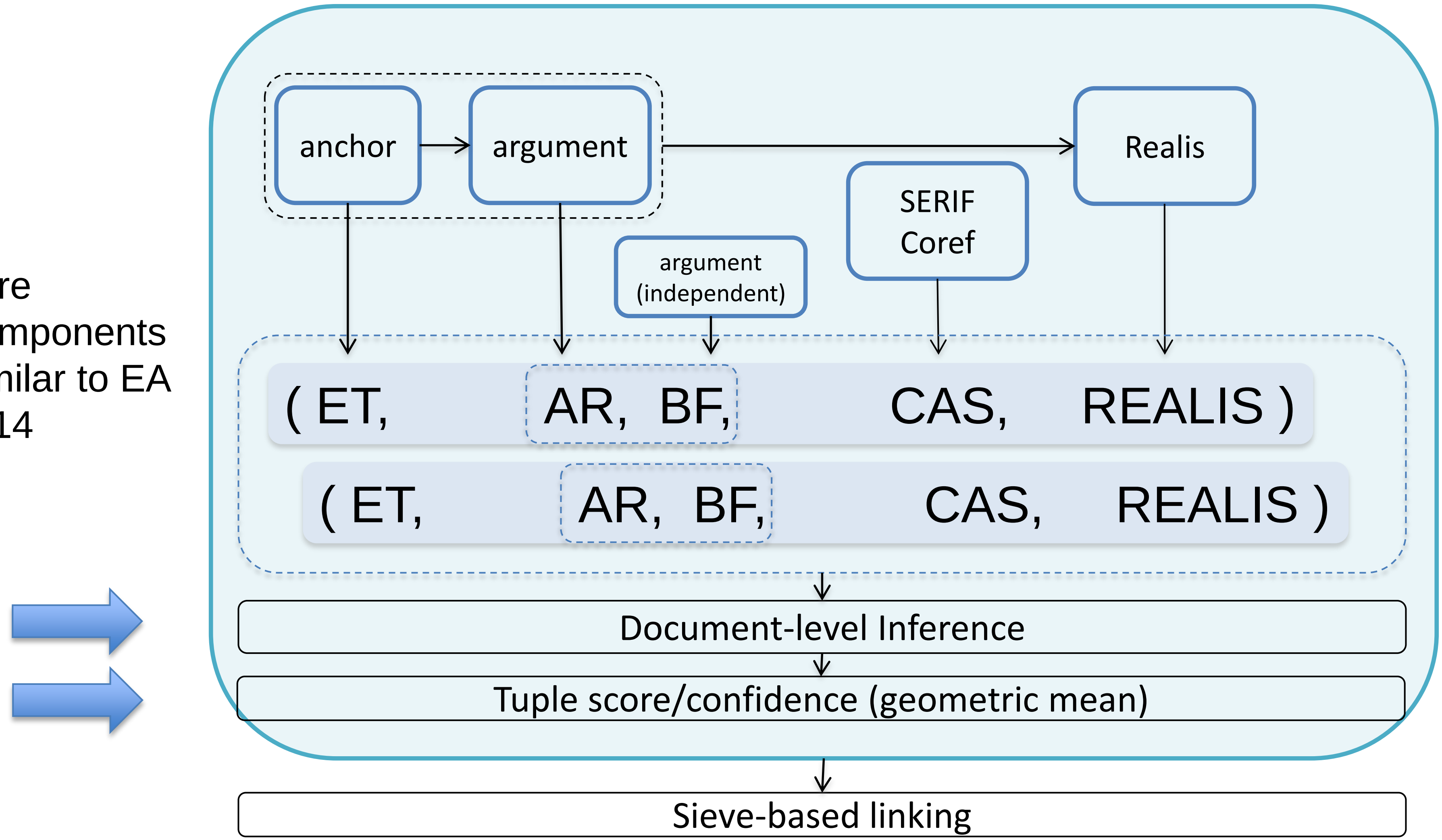
- If base filler (BF) is a Name, return the BF
- Otherwise, use coreference chains from SERIF using following rules:
  - If BF is a description:
    - When a name is available, use
      - The proper country name as appropriate
      - Otherwise, the longest name
    - Otherwise return BF
  - If BF is a pronoun:
    - When a name is available, that name (as in the description case)
    - Otherwise a description by preferring
      - Closeness in sentence distance
      - Earliness in sentence
      - Length

Realis Model

- Predict types (Actual, Generic, Other), based on:
  - P(asserted):
    - Syntactic-parse based rules to set P(specific) to either 0 or 1
  - P(specific):
    - Classifier trained on ACE and richERE
- Combine as:
  - P(Actual): P(asserted) * P(specific)
  - P(Generic): 1.0 – P(specific)
  - P(Other): P(specific) * (1.0 – P(asserted))

BBN System

Core Components Similar to EA 2014



Document-level Inference

- We apply the following document-level rules:
- Copy arguments between certain classes of events
  - For certain event types, when certain arguments are shared, for example
    - *Conflict.Attack* <-> *Life.Injure* when *Target/Victim* arguments are shared
  - For certain event types, when certain arguments are missing and
    - Related events with that argument appear nearby
      - *Crime* argument copied between *Justice* events
    - The independent argument model found an argument

Document-Level Inference (2)

- Location inference:
  - If *x* fills a PLACE argument role and *x* is part-of *y* (relation prediction or gazetteer), add *y* as a PLACE argument
- Delete:
  - non-GPE PLACE arguments
  - If realis is Actual, discard "*you*" base fillers
  - Discard events lacking crucial roles (e.g. *Personnel.End-Position* events lacking a *POSITION* role, *Justice.Sentence* event with no Sentence)

Final Tuple Scoring

(ET, AR, BF, CAS, REALIS)

geometric-mean($anchor_s$, $argument_s$, $coreference_s$, $realis_s$) :

- Coreference score defaults to 1.0, except when BF is non-relative pronoun then score is 0.75
- All tuples with geometric-mean larger than some threshold are kept
  - ***BBN1-BBN4:*** threshold is 0.3
  - **BBN5**: No document rescoring, target higher precisions system by only keeping tuples with high scores from independent classifiers
- Confidence in the output is the tuple score

Outline

- Overview of our system
- New for our 2015 system
  - Targeted training (trigger)
  - Embeddings (argument attachment)
  - Sieve (argument linking)
- Submissions, results, and analysis

Targeted Training

- Community resources (ACE, RichERE) serve many purposes
  - Balancing annotation density across entities, relations and events
  - Complete document annotation required for some tasks
  - Representative test set
- Achieving broad coverage training data for events with full document annotation is challenging
  - *ACE 2005*: 10 of 33 event types occur less than 25 times
  - Even when an event is common, each trigger/anchor/nugget may occur only 2 or 3 times
    - Difficult for a classifier to learn
- Explore focused training data creation for EA task
  - Data that the system can learn from (even if it is a bad test set)
  - For example, prioritize examples of the things we care about over a natural distribution

Targeted Training: Process

- Perform sentence-selected (rather than full document) annotation
- Give annotators search tools to find good sentences
  - People can intuitively think of triggers for most events
- Allow annotators to skip “confusing” examples
  - The system probably won’t be able to learn from them anyways
- Allow the annotator to mark as many words indicating an event as they would like
  - Better for the system to see multiple triggers

Targeted Training: Impact

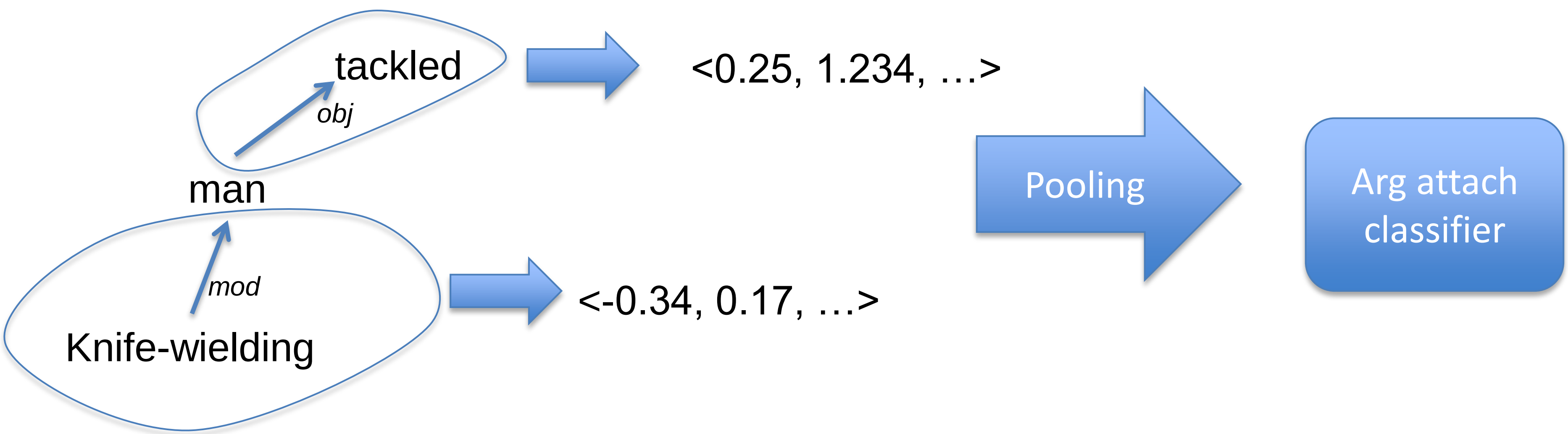
- Annotated ~5.8K positive & 6.4K negative sentences
  - Each sentence for a single event type
    - 4-8 hours per event type for all event types
    - Additional annotation for a few event types where we observed poor system performance
  - Annotation included triggers and arguments, in current system only triggers were used
    - Trigger only annotation would be faster, but include arguments for future work
  - Resulting annotation is
    - Far denser in positive examples than ACE
    - Has negative examples that are expected to be particularly useful because they are expected to be “confusable” (e.g. involve alternative senses of a potential trigger)
- BBN1 uses targeted data for training, BBN2 did not
  - ~12% relative improvement in argument score from additional training data

Context Embeddings

- Event arguments can often be distant from event triggers
- But often the argument context is informative
  – *The **knife-wielding** <u>man</u> was tackled by a bystander, but only after three people were severely injured in the <u>attack</u>.*
  – *<u>Acme Inc.</u>**’s creditors** were disappointed by Friday’s <u>bankruptcy</u> filing.*
- We would like to learn informative argument contexts which never appear in our supervised training data based on those which do

Context Embeddings: AA

- We trained dense vector representations of the normalized dependency trees contexts of words on Gigaword V5 using a variant of the skip-gram model due to (Levy & Goldberg, ‘14)
- We include this representation in our AA model



Context Embeddings: AA

- Internal development tests on KBP-2014 EA newswire eval corpus
  – Embeddings improve our 2014’s best system (BBN1), scored using 2014 EA scorer
- BBN1 used context embeddings, BBN3 did not
  – ~10% relative improvement from context embeddings

Argument Linking: Sieve

- Link arguments into event frames (EFs) using sieve-based approach
  - Applying tiers of deterministic linking decisions from highest to lowest precision
  - Sieves developed based on nw portion of the 2015 EAL training data (LDC2015E41)
    - Link-by-event-type baseline on this data was high, this informed our decisions about how to proceed
- All submissions used the same sieve-based approach
  - 4% relative improvement between sieve based approach and link-by-event type baseline

Argument Linking: Sieve

- Tiers that encourage linking
  1. Arguments that share an event anchor are group into an EF
     - This will only link arguments within a sentence
  2. Sentence internally (from left to right), merge when event frames (EF) share an event type unless
     - They violate ontology based constraints
     - We observe certain discourse connectives (e.g. cause)
  3. Across the document (from earlier to later sentences), merge unless they violate ontology based constraints
- Example ontology based constraints
  - Event role uniqueness in a single EF: Time, Place, Adjudicator (Justice), Org (Declare-Bankruptcy). Unless they are compatible: coref to same entity, compatible CAS strings, Place containment, etc.
  - Voluntary anchors should not be combined with involuntary (Crime related) anchors, e.g. ‘give’ vs ‘steal’.

Outline

- Overview of our system
- New from our 2015 system
  - Targeted training (trigger)
  - Embeddings (argument attachment)
  - Sieve (argument linking)
- Submissions, results, and analysis

System Submissions

- Submissions
  - BBN1: primary submission
  - BBN2: as BBN1 but without targeted training (trigger)
  - BBN3: as BBN1 but without embeddings (argument attachment)
  - BBN4: as BBN1 but without some inference rules added in 2015
  - BBN5: BBN1 with threshold on tuple score set high

	Targeted training (trigger)	Embeddings (AA)	2015 doc-level inference rules	Geometric mean scoring
BBN1	✓	✓	✓	✓
BBN2	×			
BBN3		×		
BBN4			×	
BBN5				×

System Submissions

- Argument linking
  - sieve heuristics based on NW portion of EAL 2015 training data (LDC2015E41)
- Training data (all submissions):
  - Anchor Classifier: ACE, richERE, targeted training (except BBN2)
  - Argument Classifier: ACE, richERE, embeddings (except BBN3)
  - Realis: ACE, richERE

Evaluation Results

	Targeted training (trigger)	Embeddings (AA)	2015 doc-level inference rules	Geometric mean scoring
BBN1	✓	✓	✓	✓
BBN2	×			
BBN3		×		
BBN4			×	
BBN5				×

	P	R	F1	EAArg	Overall
BBN1	***37***	***39***	***38***	***24***	***24***
BBN4	***37***	***39***	***38***	***24***	23
BBN2	34	37	36	21	22
BBN3	37	36	36	22	21
BBN5	***46***	29	36	21	19

- Submissions only varied in their approach to argument extraction
  - Performance improved with:
    - Targeted training (BBN1 vs BBN2)
    - Contextual embeddings (BBN1 vs BBN3)
- BBN1 is the top ranked system in the evaluation
  - BBN1's argument extraction was both higher precision and higher recall than other systems
  - Difference was much larger in recall

Analysis: Diagnostic Argument Scores

Raytheon BBN Technologies

Neutralize	EAArg BBN1	EAArg LDC	Prec BBN1	Prec LDC	Rec BBN1	Rec LDC
-	24	37	37	76	39	40
Realis	33	39	45	82	46	42
Realis & CAS	**39**	**41**	52	84	50	43

- Both assigning realis status and finding the canonical argument string negatively impact BBN performance
  - Absolute difference of either (24 -> 33 -> 39) is larger than the improvements we saw from targeted training and contextual embeddings
  - Some work in 2015 development to improve features for realis model, but still more room for improvement
- In the Neutralize Realis & CAS, BBN1 EAArg (aggregate argument) score approaches that of LDC's
  - In this setting, BBN1 exceeds LDC's recall but does not match their precision

Analysis: Linking Scores of BBN1

Raytheon
BBN Technologies

Linking Approach	EALink (nw+df)	EALink (nw)	EALink (df)
Baseline (same eventType)	22.5	21.1	24.8
System	23.4 (+0.9)	22.3 (+1.2)	25.2 (+0.4)
Max-linking	30.2 (+6.8)	30.4 (+8.1)	29.8 (+4.6)

- Some of our previously assumed unique event roles constraints are violated in the 2015 reference linking annotation:
  - Money in Transaction.Transfer-Money, Prosecutor in Justice.Trial-Hearing, Place in Conflict.Demonstrate
  - Especially true for Place of a Conflict.Demonstrate event, e.g. at 'the Washington National Cathedral', or 'around the country'.

Raytheon
BBN Technologies

Thanks!