TAC: Event Argument And Linking Evaluation Task Description (DRAFT)

March 31, 2015

Contents

Goal 2

Task 2

Differences between 2014 EA and 2015 EAL 3

Event Taxonomy 4

Marking of Realis 5

Event Hoppers 5

System Output 6

Argument System Output 6

Linking System Output 8

Offset Calculation and Formatting 8

Canonical Argument String 8

Inference and World Knowledge 9

Inferring Arguments 9

Invalid Inference of Events from Other Events 10

Invalid Inference of Events from States 10

Departures from ACE 2005 11

Corpus 13

Metadata in Source Documents 13

Assessment and Evaluation 13

Dimensions of Assessment 14

Creating Event Frames 15

Scoring 15

Data Resources for Participants 18

Submissions and Schedule 19

Submission 19
Schedule 19
Filtering of Redundant Responses during Pooling and Scoring 19
References 20

Goal

The Event Argument Extraction and Linking task at NIST TAC KBP 2015 aims to extract information about entities (and times) and the role they play in an event. The extracted information should be suitable as input to a knowledge base. Systems will extract event argument information that includes *(EventType, Role, Argument)*. The arguments that appear in the same event will be linked to each other. EventType and Role will be drawn from an externally specified ontology. Arguments will be strings from within a document representing the canonical (most-specific) name or description of the entity.

In a KB context, this task supports queries like "List the DATE and PARTICIPANTS of Conflict.Protest LOCATED in Springfield". This task is an extension of the 2014 Event Argument Extraction task and will be evaluated in English only. In 2015, the linking task is limited to linking events within a document.

Task

Systems will be given a 200-500 document corpus and asked to (a) extract instances of arguments that play a role in some event and (b) group those arguments that participate in the same event. Figure 1 illustrates the input and output for a short passage. The 2015 task is illustrated in the right purple box. For reference, 2014-style output appears in the table on the left. The event-taxonomy, which specifies extractable event-types and argument roles appears in Table 1. Systems will need to identify Canonical Argument Strings(CAS), i.e. if a mention can be resolved to a name, the CAS should be the name, if the mention cannot be resolved to a name (e.g. "three police officers"), systems should return a specific nominal phrase. The linking of arguments will group arguments at the level of an event hopper. Event hoppers represent participation in what is intuitively the same event. The arguments of an event hopper must

- Have the same EventType label
- Not conflict in temporal or location scope

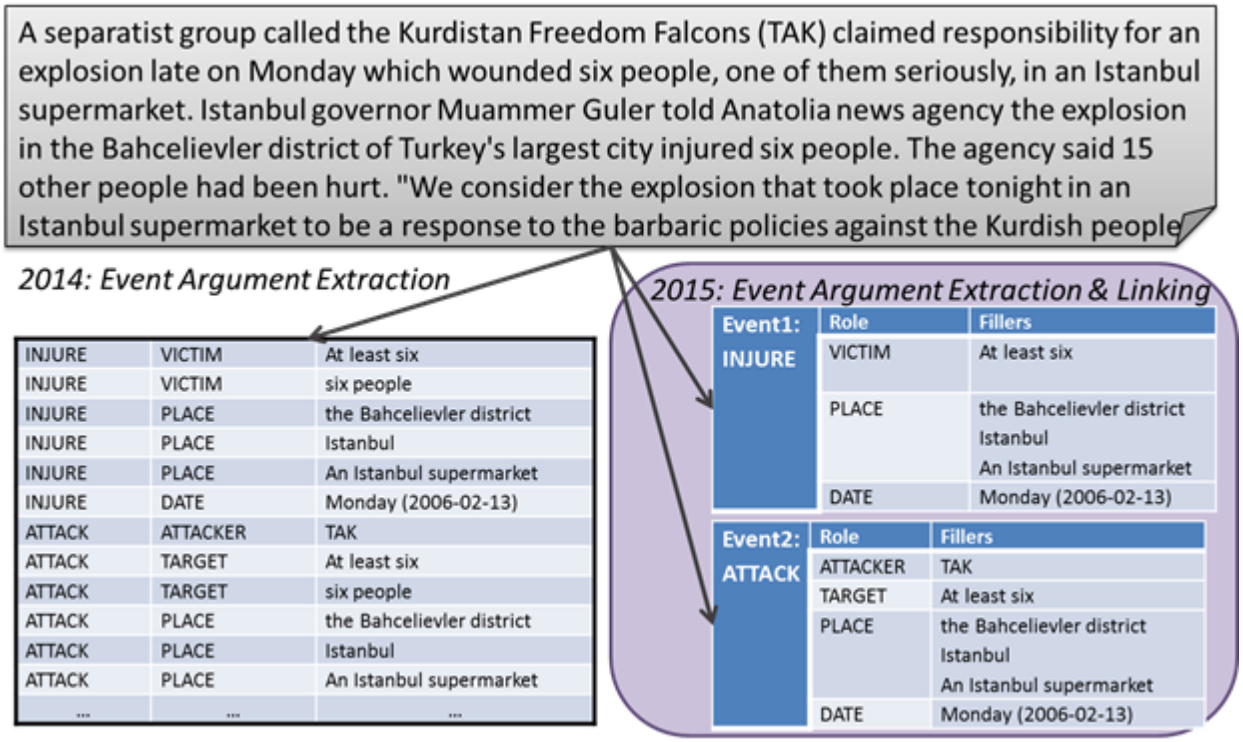

FIGURE 1: DOCUMENT TEXT, 2014 EVENT ARGUMENT TASK, AND 2015 EVENT ARGUMENT AND LINKING TASK

Differences between 2014 EA and 2015 EAL

The 2015 task:

1. Requires that arguments be grouped into Event Hoppers. The primary scoring metric will be over event hoppers. An implementation of a baseline approach to linking (link all arguments with the same event type) will be made available to all participants.
2. Adds a new event type and associated arguments: Manufacture.Artifact
3. Artifact arguments can include "Commodities". This would impact Movement.Transport, Transaction.Transfer-Ownership, and Manufacture.Artifact.

Some other changes are under discussion:

1. Modifications to the Contact and Transaction event subtypes and assessment process to include:
    a. Contact.Phone-Write renamed to includes all two (or more)-way communication that is not in person (letters, phone calls, skype, google chat, etc.)
    b. For Contact and Transaction events, assess Type at the type and subtype level. As in 2014, the official score will measure accuracy of the Type.Subtype, a lax metric that only requires producing the correct type will be calculated as a diagnostic.
2. Additional automatic filtering to reduce the number of assessments (these would be implemented in the submission/validation process)
    a. Automatic expansion of sub-sentence PJ to a predefined sentence
    b. Within a submission, automatic selection of only the highest confidence CAS in cases

where multiple tuples are identical except the CAS and the CASs are nested. For example, only selecting the first of the following tuples for assessment if all three appear in a single submission. In 2014, all were assessed, but after CAS coreference was applied typically only the most confident was scored:

i. (Life.Marry, Person, Smith, Actual, 0.9)
ii. (Life.Marry, Person, Sue, Actual, 0.5)
iii. (Life.Marry, Person, Sue Smith, Actual, 0.7);

Event Taxonomy

A system will be assessed its performance at extracting event-arguments as described in the tables below. The event and event-specific roles (argument types) are listed in Table 1. All events can also have a Time and Place argument. Only certain entity types are valid for each role. Table 2 lists the valid entity-types per-role.

Event Type	ARG1	ARG2	ARG3	ARG4	ARG5
Business.Declare-Bankruptcy	Org				
Business.Merge-Org	Org	Org			
Conflict.Attack	Attacker	Target	Instrument		
Conflict.Demonstrate	Entity[2]				
Contact.Meet	Entity[3]	Entity[3]			
Contact.Phone-Write *(may be renamed Contact.Correspondence)*	Entity[3]	Entity[3]			
Life.Marry	Person	Person			
Life.Divorce	Person	Person			
Life.Injure	Agent[1]	Victim	Instrument		
Life.Die	Agent[1]	Victim	Instrument		
Movement.Transport	Agent[1]	Artifact[1]	Vehicle	Price[2]	Origin, Destination
Personnel.Start-Position	Person	Entity[1]	Position		
Personnel.End-Position	Person	Entity[1]	Position		
Personnel.Nominate	Agent[2]	Person	Position		
Personnel.Elect	Entity[3]	Person	Position		
Transaction.Transfer-Ownership	Seller	Buyer	Beneficiary	Price[1]	Artifact[2]
Transaction.Transfer-Money	Giver	Recipient	Beneficiary	Money[1]	
Justice.Arrest-Jail	Agent[1]	Person			Crime
Justice.Release-Parole	Entity[3]	Person			Crime
Justice.Trial-Hearing	Prosecutor	Adjudicator	Defendant		Crime
Justice.Sentence	Adjudicator	Defendant	Sentence		Crime
Justice.Fine	Adjudicator	Entity[3]		Money[2]	Crime
Justice.Charge-Indict	Prosecutor	Adjudicator	Defendant		Crime
Justice.Sue	Plaintiff	Adjudicator	Defendant		Crime
Justice.Extradite	Agent[1]	Person	Origin	Destination	Crime
Justice.Acquit	Adjudicator	Defendant			Crime
Justice.Convict	Adjudicator	Defendant			Crime
Justice.Appeal	Prosecutor	Adjudicator	Defendant		Crime
Justice.Execute	Agent[1]	Person			Crime

Justice.Pardon	Adjudicator	Defendant			Crime
Manufacture.Artifact	Agent[1]	Artifact[2]	Instrument		

TABLE 1: EVENT TYPES AND ARGUMENT ROLES. TIME AND PLACE ARE VALID ARGUMENT ROLES FOR ALL EVENT TYPES. NUMBERS IN [] DISTINGUISH BETWEEN ROLE LABELS FOR WHICH THE SAME ROLE-LABEL IS ASSIGNED DIFFERENT VALID ENTITY TYPES IN TABLE 2

Role	Valid Entity Types	Role	Valid Entity Types	Role	Valid Entity Types
Adjudicator	PER, ORG, GPE	**Entity[1]**	ORG, GPE	**Position**	JOB
Agent[1]	PER, ORG, GPE	**Entity[2]**	PER, ORG	**Price[1]**	MONEY
Agent[2]	PER, ORG, GPE, FAC	**Entity[3]**	PER, ORG, GPE	**Price[2]**	NUM
Artifact[1]	PER, WEA, VEH, COM, FAC	**Giver**	PER, ORG, GPE	**Prosecutor**	PER, ORG, GPE
Artifact[2]	VEH, WEA, FAC, ORG, COM	**Instrument**	WEA, VEH	**Recipient**	PER, ORG, GPE
Attacker	PER, ORG, GPE	**Money[1]**	MONEY	**Seller**	PER, ORG, GPE
Beneficiary	PER, ORG, GPE	**Money[2]**	NUM	**Sentence**	SENTENCE
Buyer	PER, ORG, GPE	**Org**	ORG	**Target**	PER, ORG, VEH, FAC, WEA
Crime	CRIME	**Origin**	GPE, LOC, FAC	**Vehicle**	VEH
Defendant	PER, ORG, GPE	**Person**	PER	**Victim**	PER
Destination	GPE, LOC, FAC	**Plaintiff**	PER, ORG, GPE		

TABLE 2: VALID ENTITY TYPES FOR EACH ROLE.

Additional descriptions about the definition of the event types and roles can be found in LDC's assessment guidelines.

Marking of Realis

Each (EventType, Role, ArgumentString) tuple should be augmented with a marker of Realis: ACTUAL, GENERIC, or OTHER.

ACTUAL will be used when the event actually happened with the ArgumentString playing the role as reported in the tuple. For this evaluation, ACTUAL will also include those tuples that are reported/attributed to some source (e.g. *Some sources said…..*, *Joe claimed that…..*)

GENERIC will be used for (EventType, Role, ArgumentString) tuples which refer to the event/argument in general and not a specific instance (e.g. *Weapon sales to terrorists are a problem*)

OTHER will be used for (EventType, Role, ArgumentString) tuples in which either the event itself or the argument did not actually occur. This will include failed events, denied participation, future events, and conditional statements.

If either GENERIC or OTHER could apply to an event (e.g. a negated generic), GENERIC should be used.

Event Hoppers

Event hoppers represent participation in what is intuitively the same event. The arguments of an event hopper must

- Have the same EventType label
- Not conflict in temporal or location scope

An event hopper can have multiple TIME and PLACE arguments when these arguments are refinements of each other (e.g. a city and neighborhood within the city). The arguments of an event hopper need not have the same REALIS label (e.g. *John attended the meeting on Tuesday, but Sue missed it* results in a single hopper with John as an ACTUAL entity argument and Sue as an OTHER entity argument). An event hopper can have conflicting arguments when conflicting information is reported (for example conflicting reports about the VICTIM arguments of Conflict.Attack event). The same entity can appear in multiple event hoppers. Additional details about event hoppers can be found in the LDC's event argument linking guidelines [[add link]].

System Output

Submissions should be in the form of a single .zip or .tar.gz archive containing exactly two subdirectories named "arguments" and "linking", respectively. The "arguments" directory shall contain the event argument system output in the format given under "Argument System Output" below. The "linking" directory shall contain the event linking system output in the format given under "Linking System Output" below. The existence of two output files should not discourage approaches that seek to jointly perform the argument extraction and linking task.

Argument System Output

The argument output directory shall contain one file per input document (and nothing else). Each file's name should be exactly the document ID of the corresponding document, with no extension. All files should use the UTF-8 encoding.

Within each file, each response should be given on a single line using the tab-separated columns below. Completely blank lines and lines with '#' as the first character (comments) are allowable and will be ignored.

A sample argument response file can be found here:
https://drive.google.com/file/d/0BxdmkxB6KWZnV0wwcU14cFBsTjQ/edit?usp=sharing

The values in this file were automatically transformed from LDC's ACE annotation of "APW_ENG_20030408.0090". Column 7 (PJ) only includes one offset pair per response line because in ACE event extraction was limited to within sentence event-mention detection. This limitation does not hold for the TAC task. Column 9 (AJ) is NIL because argument inference in the ACE task was limited to coreference. This limitation does not hold for the TAC task.

Column #	Source	Column Name/Description	Values
1	System	**Response ID**	32-bit signed integer (-2^31 to 2^31-1), unique within each file

2	System	**DocID**	
3	System	**EventType**	From ACE Taxonomy see Table 1, column 1
4	System	**Role**	From ACE Taxonomy see Table 1
5	System	**Normalized/canonical argument string (CAS)**	String
6	System	**Offsets for the source of the CAS.**	Mention-length offset span
7	System	**Predicate Justification (PJ).** This is a list the offsets of text snippets which together establish (a) that an event of the specified type occurred, and (b) that there is some filler given in the document for the specified role. We will term the filler proven to fill this role the **base filler**. If the justifications prove there are multiple fillers (e.g. "John and Sally flew to New York"), which is to be regarded as the base filler for this response will be disambiguated by column 8. The provided justification strings should be sufficient to est9ablish (a) and (b). "Justifications which include spans not needed to establish (a) and (b) will be marked inexact. However, if the number of unnecessary supporting sentences is extreme or inconvenient for annotation, the annotators will ignore this instance (causing it to be counted wrong for purposes of scoring)." Note that the task of the predicate justification is only to establish that there is a filler for the role, not that the CAS is the filler for the role	Set of unrestricted[1] offset spans
8	System	**Base Filler (BF).** This is the base filler referred to in 7.	Mention-length offset span
9	System	**Additional Argument Justification(AJ).** If the relationship between the base filler and the CAS is identity coreference, this must be the empty set. Otherwise, this must contain as many spans (but no more) as are necessary to establish that CAS filling the role of the event may be inferred from the base filler filling the role of the event. One example of such an inference will arguments derived through member-of/part-of relations.	Set of Unrestricted offsets
10	System	**Realis Label**	{ACTUAL, GENERIC, OTHER}
11	System	**Confidence Score.** In the range [0-1], with higher being more confident. In some scoring regimes, the confidence will be used to select between redundant system responses. If necessary due to the short-assessment time frame, confidence may also be used to select	[0-1]

[1] An unrestricted offset span may always be as long as a sentence without being assessed too long, even if a shorter span is available.

		those responses to assess (e.g. assessing a system's top N responses).	

TABLE 3: COLUMNS IN SYSTEM OUTPUT

Linking System Output

The "linking" directory shall contain one file per input document (and nothing else). Each file's name should be exactly the document ID of the corresponding document, with no extension. All files should use UTF-8 encoding.

Within each file, each line will correspond to one event hopper. The line for an event hopper should contain a space-separated list of response IDs for the responses in that event hopper. These response IDs must correspond to those provided in column 1 of the files in a submission's "arguments" directory. The same response may appear in multiple event hoppers and all response for a document must appear in some event hopper, if only as a singleton. **The only exception is that any response whose realis is predicted as GENERIC by the system must not appear in the linking output.**

Completely blank lines and lines with '#' as the first character (comments) are allowable and will be ignored.

Offset Calculation and Formatting

As in TAC KBP SlotFilling, each document is represented as a UTF-8 character array and begins with the "<DOC>" tag, where the "<" character has index 0 for the document. Thus, offsets are counted before XML tags are removed. Offset spans in columns 6 to 8 are inclusive on both ends: the start offset must be the index of the first character in the corresponding string, and end offset must be the index of the last character of the string (therefore, the length of the corresponding mention string is endoffset – startoffset + 1).

Start and end offsets should be separated by a dash ("-") with no surrounding spaces and pairs of start/end offsets for different mentions should be separated by comma (",") with no surrounding spaces. For example, for the above query, if "yesterday" appears at offset 200 in the document and the document date appears at offset 20, then a valid entry for Column 5 in this case would be: 200-208,20-32 (assuming the endoffset for the document date is 32).

Canonical Argument String

Canonical Argument Strings will be one of the following:

- A string that reflects the fullest/most informative name of a PER, ORG, GPE, FAC, WEA, VEH, LOC in the document
    - Assessments will follow the TAC KBP-Slot Filling guidelines for Name Slots (section 2.1 of http://surdeanu.info/kbp2013/TAC_KBP_2013_Assessment_Guidelines_V1.3.pdf)
- A string that reflects a nominal that cannot be resolved to a name for a PER, ORG, GPE, FAC,

WEA, VEH, or LOC

- A normalized specific-date/time (in progress)
    - As in TAC KBP-SlotFilling, dates must be normalized. Systems have to normalize document text strings to standardized month, day, and/or year values, following the TIMEX2 format of yyyy-mm-dd (e.g., document text "New Year's Day 1985" would be normalized as "1985-01-01"). If a full date cannot be inferred using document text and metadata, partial date normalizations are allowed using "X" for the missing information. For example:
        - May 4th" would be normalized as "XXXX-05-04";
        - "1985" would be normalized as "1985-XX-XX";
        - "the early 1900s" would be normalized as "19XX-XX-XX" (note that there is no aspect of the normalization that captures the "early" part of the filler).
        - "the third week of June 2005" as "2005-06-XX"
        - "the third week of 2005" may be returned as **either** "2005-XX-XX" or "2005-01-XX".
- A string-fill for CRIME, SENTENCE, JOB, MONEY

Newlines and tabs in canonical argument strings

The following characters in canonical argument strings shall be replaced with a single space: Windows-style newlines ("\r\n"), Unix newlines ("\n"), and tabs ("\t").

Inference and World Knowledge

In the KBP Event Argument Extraction task, assessors will be instructed to mark an answer as correct if a reasonable reader would interpret the document as evidence that the (EventType, Role, ArgumentString, Realis) tuple is correct. They will do this even if such a judgment is derived through inference rather than, for example, a direct linguistic connection between an event-trigger and an argument. For purposes of this evaluation, systems should infer **argument participation** through links between events; however they should not infer the **occurrence** of one event from another.

Inferring Arguments

Inferences of arguments may include inferring casuality/part-of relations between the verbal-events in a passage, inferring locations through part-of relations, etc—for example inferring the Agent argument of Life.Injure from the Attacker argument of Conflict.Attack. While world-knowledge on its own is not a sufficient reason for a correct answer, such knowledge can contribute to a reasonable reader's assessment. For example, while every instance of a known terrorist group cannot be assumed to be an instance of (Conflict.Attack, Attacker), knowledge that the group has participated in terrorist activities can contribute to a reader's interpretation of vaguely worded text. In such cases, the assessor is

instructed to judge "Does this document support the claim of the (EventType, Role, NormalizedArgumentString, Realis) tuple?" Inference about geographical locations (e.g. Cambridge in Massachusetts vs. Cambridge in England) will be assessed using similar guidance.

Invalid Inference of Events from Other Events

While events can in principle be inferred from other events, for purposes of this evaluation, systems should not infer such events. This does not preclude the same text from itself justifying multiple event types (e.g. *shot* in some contexts triggers both injury and attack). This principle applies to all event types. Some particularly common examples:

- Subtypes of Life (e.g. Life.Marry from Life.Divorce)
- Subtypes of Justice (e.g. Justice.Convict from Justice.Pardon)
- Subtypes of Personnel (e.g. Personnel.Start-Position from Personnel.End-Position)

Do not infer future events from current or past events, relations or states. For example, do not infer (Life.Die, Person, Bob Smith, Other) from statements about Bob Smith's marriage, employment, etc. .

Invalid Inference of Events from States

The distinction between a stative relation and the event this relation is a consequence of can be tricky. For most events, we rely on the annotator's judgment that an event is explicitly or implicitly described in the text. The following event types require heightened scrutiny: for these, either (a) a valid temporal argument for the event to be inferred must be available or (b) the event must be signaled by textual evidence of the event (and not only the state):

- Life.Marry
- Life.Divorce
- Personnel.Start-Position
- Personnel.End-Position
- Personnel.Nominate
- Personnel.Elect
- Transport.Movement

Examples of <u>blocked</u> events

- Personnel.Start-Position
    - *ACME spokesman John Smith.*
    - *John Smith works for ACME.*
- Life.Divorce
    - *Sue and her ex-husband John share custody of their children.*
- Transport.Movement
    - *John was born in Boston and went to school in California.*

Examples of <u>allowed</u> events

- Personnel.Start-Position events
    - *ACME hired John Smith. (explicit textual description)*
    - *John Smith has worked for ACME since 2005. (DATE)*
    - *ACME's spokesman since 2005 (DATE)*

- Life.Divorce
    - *Sue, John's ex since 2000…. (DATE)*
    - *John left his wife Sue. She retained ownership of the house. (textual evidence[2])*
- Movement.Transport
    - Bob went to the airport with no particular destination in mind, and the next day he found himself in Prague. (the event is described In the text itself)
- Justice-Arrest.Jail
    - *Bob, an inmate at the county jail…* (Justice.Arrest-Jail is not on the list of event types requiring heightened scrutiny. As such, the assessor will assess this in context without heightened scrutiny).

Departures from ACE 2005[3]

While the ACE 2005 event annotation is being provided to all participants, this task diverges from ACE in some cases. One example of divergence is the addition of correct answers derived through inference/world knowledge (see above). This evaluation will treat as correct some cases that were explicitly excluded in ACE 2005.

- EventType, Role, NormalizedArgumentString tuples that a reasonable reader considers correct but are not explicitly signaled in a single sentence. Some examples are as follows, but they are by no means exhaustive:
    - Inferable arguments (e.g. Agent, Place, Time, etc.), regardless of whether they appear in sentences where ACE would have marked an event-trigger.
    - Arguments that can be inferred through implicit or explicit causality (e.g. the ATTACKER of a Conflict.Attack event also being the AGENT of Life.Die event).
        - This removes the "trumping" conditions between {ATTACK, INJURE, DIE} and {MEET, TRANSPORT, EXTRADITE}.
    - Arguments which can be inferred through implicit or explicit relations present in the document. For example, PLACE arguments can be inferred through implicit (or explicit) LocatedIn relations in the document.
- For the most part, arguments will be considered valid even independently of the other event-arguments
    - The AGENT/VEHICLE/etc. arguments of a Movement.Transport event are correct even when the ARTIFACT is unspecified (or not a WEAPON, VEHICLE or PERSON); The

[2] This is an example of context being used to interpret what could be seen as ambiguous.

[3] Light and Rich ERE annotation will also be provided to participants. The ERE definition of event and event-argument differs in some cases from both the ACE and TAC KBP definitions, but participants may still find the annotation useful. Three notable differences are: (a) ERE allows arguments outside of the sentence in which a trigger is found; (b) Light ERE does not include certain entity types (e.g. VEHICLE, WEAPON); (c) Light ERE only marks 'actual' events and not generic, future, attempted etc.

AGENT/PRICE/etc. arguments of a Transaction.Transfer Ownership is correct even when the ARTIFACT is unspecified or not a WEAPON, VEHICLE or ORGANIZATION).

  - All valid Place arguments will be considered correct (e.g. a city, state, and country). ACE only marked a single Place per 'event-mention'.

- Temporal arguments
  - Temporal arguments should be normalized using the subset of timex2 that is valid in slot-filling. (See the section on Canonical Argument Strings). Correct temporal arguments will capture a time during which the event happened/started/ended (i.e. from ACE: TIME-WITHIN, TIME-AT-BEGINNING, TIME-AT-ENDING, TIME-STARTING, TIME-ENDING, but not TIME-BEFORE or TIME-AFTER). Temporal arguments must be resolvable to a time period on the calendar (e.g. *September 2005* or the *first week of August*). Durations (*for three months*) or times marked by other events (*after his trip*) are not correct answers. Unlike ACE, we will not distinguish between different types of temporal roles, and all temporal arguments will be marked as Time.
  - In ACE, when a temporal argument might apply to multiple events, it is only marked on the most syntactically local. For this task, that restriction is removed, and temporal arguments are to be marked for all applicable events.
  - During scoring, if an assessor marks a temporal response as correct, all other response identical to that one in document Id, event type, and event role but containing less specific temporal resolutions will be deleted from both system input and the gold standard.
- Life.Injure and Life.Die. Life.Die events are frequently (perhaps always) preceded by a Life.Injure event. In ACE annotation, Life.Injure became a distinct event-mention if there was a distinct trigger "Bob was shot dead" → Life.Die and Life.Injure; "Assassins killed Bob" → only Life.Die. In this evaluation, for scoring purposes we assume Life.Die incorporates Life.Injure. If the assessed pool contains a correct Life.Die tuple, the scorer will ignore Life.Injure tuple(s) that are identical to the Life.Die tuple in CAS-id, role, and realis marker. Thus, if (Life.Die, Place, Springfield, Actual) is correct if (Life.Injure, Place, Springfield, Actual) will be ignored. This rule only applies when the Life.Die is assessed as correct. This principle may be further extended to interactions between Transaction.Transfer-Ownership and Transaction.Transfer-Money.
  - Example 2: Bob was shot and killed.
    - Correct: (Life.Die, Victim, Bob, Actual) → rule applied
    - Ignore: (Life.Injure, Victim, Bob, Actual)
  - Example 2: Bob was beheaded, but miraculously they sewed his head back on and he survived.
    - Wrong: (Life.Die, Victim, Bob, Actual) à rule not applied
    - Correct: (Life.Injure, Victim, Bob, Actual)
  - Example 3: The friendship ended when Bob brutally destroyed Joe in a game of cards.

        - Wrong: (Life.Die, Victim, Bob, Actual) à rule not applied
        - Wrong: (Life.Injure, Victim, Bob, Actual)
- This evaluation also removes three event-types from the ACE taxonomy: Life.Born, Business.Start-Org, Business.End-Org[4].

Corpus

The corpus will be a mix of newswire and discussion forum documents. The total corpus size will be 200-500 documents. A discussion forum document may contain multiple posts. The corpus will be manually and automatically filtered to ensure at least a few instances of all event-types. The discussion-forum posts will be automatically filtered to identify those posts that are not simply reposts of newswire documents. Very long discussion-forum threads will be truncated.

Metadata in Source Documents

- <DATELINE>: For newswire documents, the date in the <DATELINE> … </DATELINE> is frequently important for resolving underspecified dates (e.g. yesterday).
- <post author="..." …. >: For discussion forum data, when accurate personal pronouns (I, you) should be resolved using the string in the author attribute
- <post … datetime="2011-09-01T09:38:00" ...>: For discussion forum data, when accurate, dates should be resolved using the datetime field. Textual context can overrule the datetime field.
- <quote> … </quote>: Answers derived from <quote>...</quote> will not be assessed in this task. The pooling process will automatically remove such answers. This process will remove response rows where either the base-filler (column 8) or canonical argument string offsets (column 6) are within <quote> tags.

Assessment and Evaluation

The official score will combine a measure of extraction and grouping. System performance will be compared to a reference grouping (RG). The RG will be created by first assessing the event arguments of all submissions, then performing coreference of the arguments, and finally manually creating event hopper groupings by linking those arguments that are (a) CORRECT/INEXACT in AET, AER, BF, and CAS and (b) given an OTHER or ACTUAL realis label. The scoring metric will measure the similarity between the system's groupings and the RG.

In the event that the response pool is too large, NIST/LDC may select a subset of responses from each system to assess. For each system, tuples that are redundant in columns 2-6 and 10[5] will be collapsed to a single representative (selecting the highest confidence representative)[6]. Then the top-N answers from each system ranked by system confidence will be used to form the assessment pool.

Transformations to Submitted Output

[4] These three events are well represented as KB-SlotFilling Slots.

[5] As described in System Output, the columns labeled docid, event type, event role, CAS, CAS offsets, realis

[6] This collapsing will happen before coreference of CAS and assessment, as answers that become redundant through CAS and/or the Life.Injure à Life.Die transformation (see Departures from ACE 2005) will not be collapsed.

The following transformations may be applied to system output to minimize the assessment load:

- Responses for which the PJ and BF appear within a <QUOTE> region will be removed
- Expanding PJs to predefined sentence boundaries

Dimensions of Assessment

An LDC assessor will mark each system response for correctness. Assessment judgments evaluate the accuracy of the elements of the (EventType, Role, Cannoical Argument, Realis) tuple and of the justification for the EventType, Role, and CannonicalArgument.

Column #	Source	Column Name/Description	Values
12	Assessor	**Assessment of Event Type (AET).** Does the predicate justification (PJ) support the presence of an event of the type specified in column 3? This may be marked NIL if either the PJs or AJs for a response are excessively large. Note that a response may be marked ignore for excessively long AJ, even though the AJ normally does not play a role in evaluating AET.	C – Correct W – Wrong I - inexact –NIL
13	Assessor	**Assessment of Role (AR**)- Does the predicate justification (PJ) support the presence of who/what/when/where fills the role specified in column 4? (e.g. Can you tell from the predicate justification that some ATTACKER is mentioned for a Conflict.Attack event)? Note that this only has to prove that **some** filler is supported for this role; not necessarily the CAS (NIL if AET is WRONG or NIL)	C – Correct W – Wrong I - inexact NIL
14	Assessor	**Assessment of Normalized/Canonical Argument String**. Is the CAS a correct filler for the event-type/role tuple? (NIL if AET or AR is WRONG or NIL). Wrong (not inexact) will be used in the case of a CAS strings that are not canonical (e.g. a noun phrase when a name is available, an unnormalized date).	C – Correct W – Wrong I - inexact NIL
15	Assessor	**Assessment of base filler.** Is the base filler a correct filler for the event-type/role specified in columns 3 and 4? NIL if AET and/or AR are WRONG or NIL	C – Correct W – Wrong I - inexact NIL
16	Assessor	Coreference IDs for CAS. Singletons should be also be assigned coreference IDs. NIL if 14 is NIL or Wrong.	Unique ID (32-bit signed integer) NIL
17	Assessor	Realis Label	{ACTUAL, GENERIC, OTHER} NIL if AET or AR is

			NIL or WRONG
18	Assessor	Mention-type for CAS	{NAME, NOMINAL} NIL if AET or AR is NIL or WRONG

TABLE 4: QUESTIONS OF LDC ASSESSOR TASK

Creating Event Frames

System performance will be evaluated by comparing system output on each document to reference event frames (REFs). REFs will be created in the following way:

1. All responses from all event frames from all systems will be pooled.
2. As in the 2014 task, LDC assessors will
   a. assess all responses in this pool as described in "Dimensions of Assessment" above.
   b. group the canonical argument strings from all responses into coreference clusters
3. A *linking response pool* will be formed from all responses which
   a. are CORRECT/INEXACT in AET, AER, BF, and CAS, and
   b. have a realis label of OTHER or ACTUAL[7]
4. All responses in the response pool will be automatically grouped into equivalence classes called TRFRs[8] based on event type, event role, realis, and CAS coreference cluster. The set of these TRFRs will be referred to as the *TRFR pool*.
5. LDC annotators will group all TRFRs in the TRFR pool into the reference event frames.

Scoring

A package to automatically validate system output and score (given assessments) is available here: https://github.com/rgabbard-bbn/kbp-2014-event-arguments . It will be extended to calculate the event argument linking metric in addition to the argument only score. Systems which are written in JVM-based languages are encouraged to use the classes here directly for representing and writing their output.

Official Metric: Event Argument Linking

Event Argument Extraction Sub-score (unnormalized)

For the event argument extraction sub-score we use the linear function $TP_{EAE} - \beta FP_{EAE}$ for some parameter β, where TRFR true and false positives are defined as in the 2014 KBP EA Standard metric. Intuitively, this corresponds to a model where the user of an EAEL system derives utility 1 from a TP_{EAE} and loses utility $\beta$from an FP_{EAE}. Note that this differs from the F-measure-based score used in 2014. We will continue to report the F-based metric as an independent diagnostic measure of extraction performance (ignoring linking)

[7] As a result of this, recall of GENERIC event arguments will only be reported as a part of the diagnostic 2014 metric, not as a part of the official Event Argument and Linking metric.
[8] for 'Type, Role, [Normalized] Filler, Realis"

Linking Sub-score (unnormalized)

There are a number of clustering metrics available, including CEAF, B^3, BLANC, etc. Many of them can be straightforwardly applied to event frames subject to the modification that TRFRs may appear in multiple frames.

We propose to use the following variant of B^3[9]:

1. Let $L\ \ (d)$ be the system-provided TRFR linking for a document d. Let $R(d)$ be the reference TRFR linking, where the *i*th event frame is a set of TRFRs denoted $R_i(d)$. Define $\widehat{L\ \ (d)}$to be $L\ \ (d)$ with all TRFRs not found in $R\ \ (d)$ removed (that is, $L\ \ (d)$ without EAE false positives).
2. Define $\nu_Y(x)$ for a linking $Y$to be $(\bigcup_{Z\ \in Y s.t. x \in Z} Z) - x$ (that is, all TRFRs which are present in a common event frame with x, excluding $x$itself).
3. Define $f_{Y,Z}(x)$, the per-TRFR link F-measure, as:
    a. If $x$is not in Z, $f(x) = 0$
    b. If $x\ \in Z$and $\nu_Y(x)$ and $\nu_Z(x)$ are empty, then $f(x) = 1$.
    c. Otherwise, let $p_{Y,Z}(x)$, the precision, be $\frac{|\nu_Z(x) \cap \nu_Y(x)|}{|\nu_Z(x)|}$. Let $r_{Y,Z}(x)$, the recall, be $\frac{|\nu_Z(x) \cap \nu_Y(x)|}{|\nu_Y(x)|}$. $f_{Y,Z}(x) = \frac{2 p_{Y,Z}(x) r_{Y,Z}(x)}{p_{Y,Z}(x) + r_{Y,Z}(x)}$
4. Let $U_X(d)$ be the union of all event frames in X. We define $S_{EAL}(d, R, L)$as $\sum_{x \in U_R(d)}\ \ f_{\widehat{L,R}}(x)$. Intuitively, it is the sum of the link F scores for each TRFR present in the gold standard.

Aggregating S_{EAE} and S_{EL}

We aggregate the scores as $S_{EAEL} = \lambda[max(0, S_{EAE})] + (1 - \lambda) S_{EAL}$. Note that while S_{EAE}can be negative, we clip it to 0 on a per-document basis. If we wish to get a document-level normalized score for diagnostic purposes, we can divide by the maximum possible utility n, where n is the number of TRFRs in the reference linking. Similarly, for diagnostic purposes we can compute document-level $S_{EAE}$and S_{EAL}nscores by dividing the raw scores by n.

Aggregating Scores across the Evaluation Corpus

We define the score of a corpus as a $\sum_{d \in D}\ \ S_{EAEL}(d)/N$where D is the set of documents and N is the total number of correct TRFRs across the corpus.

Official Ranking Score

For the official ranking score, we will use $\beta = 1/4, \lambda = 1/2$ to weigh argument extraction and linking performance roughly equally[10] and to encourage high recall while maintaining reasonable precision.

---

[9] While B^3 has fallen out of favor for coreference evaluations due to its tendency to compress scores into a small range when there are many singletons, singletons are far less common in the EAEL task, so this does not appear to be a concern. In ACE annotation, event frame sizes of two and three are most common and are twice as likely as singletons.

[10] This is not exact because the ranges of likely variation of the two sub-scores differ somewhat and recall affects linking scores because you can't link what you can't find.

Because the choice of these parameters is somewhat arbitrary and has a significant impact on the evaluation, we are open to input from participants about what they should be. We will also do an analysis of the sensitivity of the final ranking to variation in the parameters.

The score used for final system ranking will be the median corpus-aggregated S_{EAEL} over 1,000 corpora bootstrap-sampled from the LDC-provided evaluation corpus. We will report for each rank the fraction of samples on which it outperforms each other rank.

Diagnostic Metric: Argument Only Scoring

As a diagnostic, we will report argument only scores for a systems. The argument only score will be the same as the score used in the TAC 2014 Event Argument Evaluation and is calculated as follows.

- Recall that certain responses will be removed from both the system output and the annotation pool before scoring. For example, see the treatment of Life.Injure in "Departures from ACE".
- In all cases below, a 'response' is defined as columns 2-8 and 10 of the system output.
- Confidence is not considered part of the response, but confidences are used to define a confidence function *c(x)*.
- Recall the CAS stands for *canonical argument string* (column 5 of the system response).
- Recall that annotators will indicate coreference relationships between CASes within the same document (column 16 of the annotation). The transitive closure of this coreference relation defines an equivalence relation which we will call **C**.
- of the following are true
    - $\text{xdocument id}=\text{ydocument id}$
    - $\text{xevent type}=\text{yevent type}$
    - $\text{xargument role}=\text{yargument role}$
    - $\text{xcas}\sim\text{Cycas}$
    - $\text{xrealis}=\text{yrealis}$
- Let **O** be the set of system responses. Define **O'** to be $\{\rho(X)|X\in O/\sim T\}$, where $\rho(X)$ is $\text{argmax}\in X\ c(x)$. If this maximum is not unique, ties will be broken 'randomly'.
- Let **P** be the union of the system responses of all systems, including the human 'system's'. Note that by definition $O'\subseteq P$.
- Let ***α*** be a function from system responses to their annotation tuples.
    - ***α*** will be completely defined on **P**.
    - Note that since ***α*** is a function, we assume there are not multiple inconsistent annotations for any response.
- We define an annotation tuple α to be a *good annotation for* a response *x* if none of columns 11-15 of α are WRONG or IGNORE and $\alpha\text{realis}=\text{xrealis}$.
    - Informally, this means the response is correct except for possible errors regarding the extent of justifications.
- We define an annotation tuple α to be a *perfect annotation for* a response *x* if columns 11-15 of α are CORRECT and $\alpha\text{realis}=\text{xrealis}$.
    - Informally, this means the response is exactly correct.
- We define an equivalence class **X** to be *good* if there exists an $x\in X$ such that $\alpha(x)$ is a good annotation for *x*.
- We define an equivalence class **X** to be *perfect* if there exists an $x\in X$ such that $\alpha(x)$ is a perfect annotation for *x*.

- We define an equivalence class **X** to be *semantically correct* with respect to **P** if **X** is non-empty and there exists $p \in P$ such that *p* is equivalent to the elements of **X** under **T** and $\alpha(p)$ is a good annotation for *p*.
    - Informally, this means we have evidence somewhere in the pool that the corresponding *docid-type-role-CAS-realis* tuple is correct, but there may or may not be a matching justification in the system output.
- Let $s \in \{\text{good, perfect, semantically correct}\}$. Define the *s-precision* of a system's output **O'** as the number of *s* equivalence classes of **O'** under **T** divided by the total number of equivalence classes of **O'** under **T.**
- Let $s \in \{\text{good, perfect, semantically correct}\}$. Define the *s-recall* of a system's output **O'** as the number of equivalence classes of **O'** under **T** which have property *s* divided by the total number of *s* equivalence classes of **P** under **T.**
    - Note the stipulation "which also has property *s"* is necessary because the responses which have a good or perfect annotation in an equivalence class of **P** may be missing from the matching equivalence class in **O'**.
- Let $s \in \{\text{good, perfect, semantically correct}\}$. Define the *s-F1* measure of a system's output **O** as $2prp+r$ where *p* is the *s*-precision of **O'** and *r* is the *s*-recall of **O'**.
- The first scoring measure for the task is good-F1, called the **standard** measure.
- The second scoring measure for the tasks is perfect-F1, called the **strict** measure.
- The third scoring measure for the tasks is semantically-correct-F1, called the **lax** measure.

Additional Diagnostic Metrics

The following diagnostic measures will be calculated but will not be used for system ranking:

- scores for newswire only and discussion forums only
- scores on the LDC-provided evaluation corpus only, without sampling
- a 2014 KBP EA-style argument extraction score (F1 over the assessment pool)
- graphs of how systems' evaluation scores would vary with changes to β and λ.
- a "macro-F" version of the score, where we compute the score on a per-document basis and take the mean.

Data Resources for Participants

Participants will have the opportunity to request the following pre-existing resources from LDC. While these resources diverge from the EA-linking task in some dimensions, they still provided useful training data for many 2014 EA systems.

- ACE 2005 Multilingual Training Data (LDC2006T06)
- DEFT ERE Data (LDC2014E31)
- Rich ERE Training Data (on-going releases during development period)
- Event Nugget and Event Nugget Coreference training data (on-going releases during development period)

Participants will also be provided with the assessments from the 2014 Event Argument Task. LDC will provide sample linking for ~50 files from the 2014 assessments. This linked data will most closely mirror the 2015 evaluation task- it will not however include the new event/role types.

Submissions and Schedule

Submission

Systems will have up to one week to process the evaluation documents. Submissions should be fully automatic and no changes should be made to the system once evaluation corpus has been downloaded. Up to five alternative system runs may be submitted per-team. Submitted runs should be ranked according to their expected overall score. Teams should submit at least one version of their system that does not access the web during evaluation. Any web-access of alternative systems should be documented in the system description.

Schedule

February 2015	Task definition released. Pre-existing resources (ACE data, ERE data, 2014 EA data) available to participants as they sign up
April 1, 2015	LDC releases sample documents (~50) with linking annotation.
April 15, 2015	Draft/beta versions of submission format and software for validation, baseline linking and scoring released
June 1, 2015	Final versions of guidelines and software released
August 2015	Evaluation Period

Filtering of Redundant Responses during Pooling and Scoring

Text:

- s6: The trial started early in the morning and lasted for six hours before the ruling was announced by Judge Jones Chinyama.
- s7: Jones Chinyama said the prosecution team had failed to prove the case against Chiluba.
- s8: The judge said that an important attestor failed to show up during the trial, adding that "I find that the accused is not guilty on all counts".

(A) Responses from System X

- R1: (Justice.Trial-Hearing, Adjudicator, Jones Chinyama, ACTUAL) CAS offsets: S6; Base Filler: Jones Chinyama, S6; conf: 0.9
- R2: (Justice.Trial-Hearing, Adjudicator, Jones Chinyama, ACTUAL) CAS offsets: S6; Base Filler: Jones Chinyama, S7; conf: 0.8
- R3: (Justice.Trial-Hearing, Adjudicator, Jones Chinyama, ACTUAL) CAS offsets: S7; Base Filler: Jones Chinyama, S7; conf: 0.7
- R4: (Justice.Trial-Hearing, Adjudicator, Jones Chinyama, ACTUAL) CAS offsets: S6; Base Filler: The judge, S8; conf: 0.6
- R5: (Justice.Trial-Hearing, Adjudicator, Jones Chinyama, ACTUAL) CAS offsets: S7; Base Filler: The judge, S8; conf: 0.5

(B) Automatically clustered responses from System X

- (R1 conf:0.9, R2 conf: 0.8, R4 conf: 0.6)
- (R3 conf: 0.7, R5 conf: 0.5)

(C) Responses from System X that are sent to assessment

- R1: (Justice.Trial-Hearing, Adjudicator, Jones Chinyama, ACTUAL) CAS offsets: S6; Base Filler: Jones Chinyama, S6; conf: 0.9
- R3: (Justice.Trial-Hearing, Adjudicator, Jones Chinyama, ACTUAL) CAS offsets: S7; Base Filler: Jones Chinyama, S7; conf: 0.7

(D) Assessment will add coreference IDs to CAS,leading to redundant responses:

(R1 conf: 0.9, R3 conf 0.7)

(E) Scoring will use R1 and ignore R3

References

ACE 2005 Training Data: http://catalog.ldc.upenn.edu/LDC2006T06

ACE Task Guidelines: https://www.ldc.upenn.edu/collaborations/past-projects/acehttp://catalog.ldc.upenn.edu/LDC2006T06

TAC 2014 Event Argument Task Definition: http://www.nist.gov/tac/2014/KBP/Event/guidelines/EventArgumentTaskDescription.09042014.pdf

TAC 2014 Event Argument Assessment Guidelines: http://www.nist.gov/tac/2014/KBP/Event/guidelines/TAC_KBP_2014_Event_Argument_Extraction_Assessment_Guidelines_V1.3.pdf

http://catalog.ldc.upenn.edu/LDC2006T06