TAC 2016: CROSS-CORPUS EVENT ARGUMENT AND LINKING EVALUATION
TASK DESCRIPTION (DRAFT)

JUNE 20, 2016

Contents

TAC 2016 Event Argument and Linking Task Description: **DRAFT**
This submission does not contain any material which may be restricted under EAR/ITAR

1   Goal

The Event Argument Extraction and Linking task at NIST TAC KBP 2016 aims to extract information about entities (and values) and the roles they play in events. 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 2016, the task introduces a cross-corpus component. In addition to linking the arguments that play some role in the same event within a single document, the participants are asked to provide a global ID to each document level event-frame.

2   Task

Systems will be given a ~90K document corpus consisting of Spanish, Chinese, and English documents. The corpus will be roughly evenly divided between the three languages and genre. Participants will be asked to:

1. For each document, extract instances of arguments that play a role in some event (same as 2014 and 2015)
2. For each document, group those arguments that participate in the same event to create a set of event frames (same as 2015)
3. Group the document-level event frames that represent the same event to create a set of corpus-level event frames (new for 2016).

Figure 1 illustrates the inputs and outputs for three English passages and one event type (CONTACT.MEET).
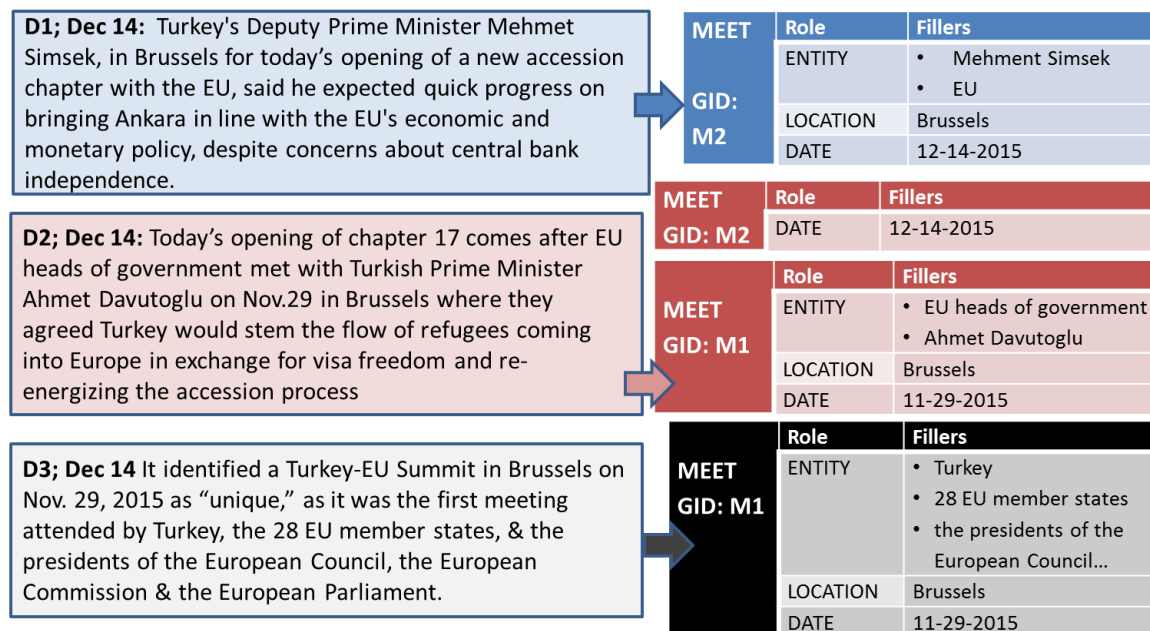
TAC 2016 Event Argument and Linking Task Description: **DRAFT**
This submission does not contain any material which may be restricted under EAR/ITAR

Figure 1: TAC Cross-Corpus EA Task

As in the 2014 and 2015 EA tasks, 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. Cross-document event coreference (as indicated using global IDs) will also operate at the level of the event hopper. The arguments of an event hopper must

- Have compatible Event Types and subtypes (identical types are always compatible; certain other types might be allowed, see below) Not conflict in temporal or location scope

2.1 Differences between 2015 EAL and the 2016 Cross-Corpus EAL task

There are a few differences between the 2015 and 2016 task:

1. As described above, the 2016 task will include Chinese and Spanish documents (in addition to English documents).
   a. Note: We will provide diagnostic scores that report performance over each language independently to allow participation in only one (or two) of the languages
2. As described above, the 2016 task will include a cross-corpus component and operate with a much larger corpus
   a. Note: We will provide diagnostic scores that report document-level metrics for participants who only want to participate in the within document task.
   b. Note: If there is sufficient interest, we can offer a post-evaluation window for the within document task only that operates over ~500 rather than 90K documents.

3. The event taxonomy will be reduced (see Table 1**Error! Reference source not found.** for the event types that will be evaluated in TAC 2016)
4. The within-document ARG and DOC-LINK scores will be calculated using a RichERE Gold Standard and not assessments (see Section 7.1)
   a. Note: LDC will perform QC and augmentation over the gold standard used in this evaluation to try to ensure the implicit/inferred arguments from the 2014 and 2015 evaluation are still incorporated into the gold standard. However, we expect there to be some cases where an assessor would judge something to be correct, but an annotator creating gold standard will miss the arguments. The organizers would be interested in learning about such instances from participants.
   b. Note: The move to a gold standard will require that system argument extents (names, nominals) be aligned with RichERE annotation. We have designed the alignment process to be generous, it may be necessary to introduce additional constraints (e.g. requiring some threshold of character overlap) if submissions are overly aggressive in making use of the generosity. Our aim with any such change will be to (a) avoid penalization of minor extent errors and/or opinions about what is correct, (b) expect that systems in general should provide reasonable names or base NPs as arguments, (c) assume that scoring of identifying the "correct" extent of a NP/name is the domain of the EDL evaluation and not the event argument evaluation.
5. A minor change to the format of the linking output at the document level to support the corpus level output (see Section 3.2)

2.2 Event Taxonomy

A system will be scored for 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. All events except Movement.Transport-Person and Movement.Transport-Artifact can have a Place argument. For the movement events, the taxonomy requires systems to distinguish between Origin and Destination. Additional descriptions about the definition of the event types and roles can be found in LDC's RichERE guidelines. For the EAL task, systems are asked to combine the RichERE event type and subtype in a single column in the system output by concatenating the type, a ".", and the subtype (see column 1 of Table 1).

EAL Event Label (Type.Subtype)	Role	Allowable ARG Entity/Filler Type
Conflict.Attack	Attacker	PER, ORG, GPE
	Instrument	WEA, VEH, COM
	Target	PER, GPE, ORG, VEH, FAC, WEA, COM
Conflict.Demonstrate	Entity	PER, ORG
Contact.Broadcast *(*this may be filtered before scoring*)	Audience	PER, ORG, GPE
	Entity	PER, ORG, GPE
Contact.Contact *(*this may be filtered before scoring*)	Entity	PER, ORG, GPE

Contact.Correspondence	Entity	PER, ORG, GPE
Contact.Meet	Entity	PER, ORG, GPE
Justice.Arrest-Jail	Agent	PER, ORG, GPE
	CRIME	CRIME
	Person	PER
Life.Die	Agent	PER, ORG, GPE
	Instrument	WEA, VEH, COM
	Victim	PER
Life.Injure	Agent	PER, ORG, GPE
	Instrument	WEA, VEH, COM
	Victim	PER
Manufacture.Artifact	Agent	PER, ORG, GPE
	Artifact	VEH, WEA, FAC, COM
	Instrument	WEA, VEH, COM
Movement.Transport-Artifact	Agent	PER, ORG, GPE
	Artifact	WEA, VEH, FAC, COM
	Destination	GPE, LOC, FAC
	Instrument	VEH, WEA
	Origin	GPE, LOC, FAC
Movement.Transport-Person	Agent	PER, ORG, GPE
	Destination	GPE, LOC, FAC
	Instrument	VEH, WEA
	Origin	GPE, LOC, FAC
	Person	PER
Personnel.Elect	Agent	PER, ORG, GPE
	Person	PER
	Position	Title
Personnel.End-Position	Entity	ORG, GPE
	Person	PER
	Position	Title
Personnel.Start-Position	Entity	ORG, GPE
	Person	PER
	Position	Title
Transaction.Transaction (*this may be filtered before scoring)	Beneficiary	PER, ORG, GPE
	Giver	PER, ORG, GPE
	Recipient	PER, ORG, GPE
Transaction.Transfer-Money	Beneficiary	PER, ORG, GPE
	Giver	PER, ORG, GPE
	Money	MONEY
	Recipient	PER, ORG, GPE
Transaction.Transfer-Ownership	Beneficiary	PER, ORG, GPE

	Giver	PER, ORG, GPE
	Recipient	PER, ORG, GPE
	Thing	VEH, WEA, FAC, ORG,COM

Table 1: Valid Event Types, Subtypes, Associated Roles for TAC 2016 EAL. The last column provides the valid Rich ERE entity type/filler type for an argument with the specified role. This column is provided to help participants understand the taxonomy. In the 2016 EAL task, participants are not required to report entity types. All events can also have a TIME role. All events except Movement.* events can also have a PLACE role.

2.3 Marking of Realis

Each (EventType, Role, ArgumentString) tuple should be augmented with a marker of Realis: ACTUAL, GENERIC, or OTHER. Complete annotation guidelines for Realis can be found in the RichERE guidelines. To summarize, ACTUAL will be used when the event is reported as actually having 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.

The scoring process automatically maps ERE annotation to argument-level realises by the following rules:

- If the ERE event mention has generic realis, all its argument will have realis GENERIC
- Otherwise,
  - If the argument's realis is marked in ERE as IRREALIS , the KBP EAL realis will be OTHER
  - Otherwise, the KBP EAL realis will be ACTUAL

2.4 Event Hoppers

Event hoppers are the unit of event coreference defined for RichERE. Full annotation guidelines with examples appear in LDC's Rich ERE annotation guidelines. To summarize, event hoppers represent participation in what is intuitively the same event. The arguments of an event hopper must

- Conceptually, be a part of the same class in the event ontology
  - The EAL submission format merges RichERE event type and subtype.
  - For most event subtypes, both the type and subtype must be the same for RichERE to consider the event mentions a part of the same event hopper
  - In LDC Rich ERE event mention annotation, Contact.Contact and Transaction.Transaction

TAC 2016 Event Argument and Linking Task Description: **DRAFT**
This submission does not contain any material which may be restricted under EAR/ITAR

are used when the local context is insufficient for assigning a more fine grained subtype. During the event hopper creation process events with Contact/Transaction subtypes may be merged with event mentions with a more specific event subtype, for example in Rich ERE a Conctact.Contact event mention can occur in the same event hopper as a Contact.Meet event.

● 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.

3   System Output

Submissions should be in the form of a single .zip or .tar.gz archive containing exactly three subdirectories named "arguments" , "linking", and "corpusLinking", 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 document-level event linking system output in the format given under "Document-Level Linking System Output" below. The "corpusLinking" directory shall contain the corpus-level event-linking output in the format given under "Corpus-Level Linking System Output" below. The existence of three outputs should not discourage approaches that seek to jointly perform the argument extraction, document-level linking, and corpus-level linking tasks.

3.1   Argument System Output → ./arguments

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 must use 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[1]:

https://drive.google.com/file/d/0Bxdmkxb6KWZnV0wwcU14cFBsTjQ/edit?usp=sharing

---

[1] 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**	A string ID, containing no whitespace, unique within a document. Such IDs may be generated using the provided Java API or by any other means a participant choses.
2	System	**DocID**	The LDC document ID
3	System	**EventType.Subtype**	From Event Taxonomy see Table 1 column 1
4	System	**Role**	From Event Taxonomy see Table 1 column 2
5	System	**Normalized/canonical argument string (CAS)**	String with normalizations (see below)
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 establish (a) and (b). 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 offset spans. No more than three offset spans may be supplied and each offset span may be at most 200 characters.
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	Set of Unrestricted offsets

8

		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 is arguments derived through member-of/part-of relations.	
10	System	**Realis Label**	One of { 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	[0-1]

3.2 Document-Level Linking System Output → ./linking

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 must use UTF-8 encoding.

Within each file, each line will correspond to one event hopper. The line for an event hopper should contain exactly two tab-separated fields. The first shall contain an arbitrary event hopper ID which must be unique within a document. No hopper ID shall contain a "-" character. The second shall 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 responses 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. **Note that this format differs from the 2015 linking format by adding event hopper IDs.**

3.3 Corpus Event Hoppers → ./corpusLinking

The "corpusLinking" directory shall contain a single UTF-8 encoded file, "corpusLinking". Within this file, each line will correspond to one corpus-level event hopper. The line for an event hopper shall contain two tab separated-fields. The first field shall be an arbitrary unique corpus-level event hopper ID. The second field shall contain a space-separated list of document-level event hoppers. Each event hopper shall be represented as the concatenation of three strings: the document ID, a "-" character, and the document-level hopper ID. Completely blank lines and lines with '#' as the first character (comments) are allowable and will be ignored.

3.4 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

9

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 end offset for the document date is 32). Participants are encouraged to use the scoring software to ensure that their system produces valid offsets.

Note: This offset definition is the same across all NIST 2016 TAC KBP tasks and the same as the 2014, 2015 EAL task. It differs from the offset definition used in LDC's ACE and Rich ERE. The README in each LDC package provides the official definition of offsets in that language.

3.5 Canonical Argument String

Canonical Argument Strings will be one of the following:

- A valid name extent as defined by the definition of names in the RichERE annotation guidelines. The automatic scoring software will align system responses with entities and treat as Correct any mention marked as NAME but as False Positive a non-named mention in cases where the Rich ERE entity includes at least one name.

- 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

    o 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".

10

- A string-fill for CRIME, SENTENCE, JOB, MONEY

3.5.1 Newlines and tabs in canonical argument strings

4 The following characters in canonical argument strings shall be replaced with a single space before submission: Windows-style newlines ("\r\n"), Unix newlines ("\n"), and tabs ("\t"). Normalization applies only to CAS strings output by system submissions; offsets should correspond to original input document text without any normalization applied.Inference and World Knowledge

In the KBP Event Argument Extraction task, systems should return all (EventType, Role, ArgumentString, Realis) tuples that a reasonable reader would understand from a document event if such understanding is derived through inference rather than, for example, a direct linguistic connection between an event-trigger and an argument. As is true in Rich ERE, an argument can be correct even if it does not appear in the same sentence as an event trigger. The gold standard for this evaluation will undergo a QC-pass designed to ensure that certain classes arguments that may have been missed (or "trumped") in annotation guidelines are correct and that cross-sentence arguments that are inferred through event causality can be included (e.g. the location of a Life.Die event being transferred to the location of Conflict.Attack event).

4.1 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)[2]. This principle applies to all event types.

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.

4.2 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):

- Movement.Transport-*

Examples of <u>blocked</u> events

---

[2] See LDC's RichERE guidelines double tagging examples for cases where a single word/phrase indicates multiple events.

11

- Movement.Transport-Person
  - *John was born in Boston and went to school in California.*

Examples of <u>allowed</u> events

- Movement.Transport-Person
  - 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)

5 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.

- RichERE allows for double tagging. See the RichERE guidelines for a discussion of double tagging.

- 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) LOCATED-IN 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 AGENT/PRICE/etc.

---

[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.

arguments of a TRANSACTION.TRANSFER-OWNERSHIP is correct even when the ARTIFACT is unspecified or not a WEAPON, VEHICLE or ORGANIZATION).

- o   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

  - o   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.

  - o   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.

  - o   If a temporal TRFR is 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.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 ERE annotation 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, (LIFE.INJURE, PLACE, Springfield, ACTUAL) will be ignored. This principle may be further extended to interactions between Transaction.Transfer-Ownership and Transaction.Transfer-Money.

  - o   Example 1: *Bob was shot and killed*.

    - ▪   Correct: (LIFE.DIE, VICTIM, Bob, ACTUAL) → rule applied
    - ▪   Ignore: (LIFE.INJURE, VICTIM, Bob, ACTUAL)

  - o   Example 2: *Bob was beheaded, but miraculously they sewed his head back on and he survived*.

    - ▪   Wrong: (LIFE.DIE, VICTIM, Bob, ACTUAL) → rule not applied
    - ▪   Correct: (LIFE.INJURE, VICTIM, Bob, ACTUAL)

  - o   Example 3: *The friendship ended when Bob brutally destroyed Joe in a game of cards*.

    - ▪   Wrong: (LIFE.DIE, VICTIM, Bob, ACTUAL) → rule not applied
    - ▪   Wrong: (LIFE.INJURE, VICTIM, Bob, ACTUAL)

TAC 2016 Event Argument and Linking Task Description: **DRAFT**
This submission does not contain any material which may be restricted under EAR/ITAR

6 Corpus

The corpus will be a mix of newswire and discussion forum documents. The total corpus size will be ~90K documents. The corpus will be a mix of Spanish, Chinese, and English. The 2016 EAL task will share the same corpus with the 2016 ColdStart and EDL tasks.

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.

6.1 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 possible personal pronouns (I, you) should be resolved using the string in the author attributes. Per RichERE standards, post authors are considered named entities and are valid event arguments and query entry points.
- <post … datetime="2011-09-01T09:38:00" ...>: For discussion forum data, when possible, 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 scored in gold-standard based metrics (the arg and doc-link subscores). The scoring 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.

7 Gold Standard Alignment and Evaluation

Scoring will have two-components: a document-level score over a RichERE annotated subset of the corpus and corpus-level score over a set of queries.

7.1 Document-level scoring

ERE annotation will be translated into a set of (*docID, event type, argument role, entity ID, realis*) TRFR tuples by straightforwardly extracting this information from every event mention argument. An exception is that for temporal arguments the entity ID will instead be the TIMEX resolution and non-temporal value arguments will use their offsets as their ID

Doing a similar translation for system responses is also straightforward except for the entity ID. To compute determine the entity ID for a system response, all documents will be parsed using Stanford's CoreNLP. For each system response, its head will be found according to Collins-style head rules and attempt to match it against the mention heads in the ERE annotation. The heads will be judged to match if there is an exact match of offsets or if there ERE head's offsets contain the systems' head offsets. If there is a match, the ERE entity ID of that mention will be used. Otherwise, a unique entity ID will be generated; this entity ID will be shared by all other arguments from that system with the same CAS. Temporal and value arguments will be special-cased as above. This alignment algorithm will likely

14

be tweaked before the final version of the guidelines to be more generous to systems.

Once both the gold standard and system responses have been translated into tuples, calculation of true positives (TP_{EAE}), false positives (FP_{EAE}), and false negatives (FP_{EAE}) is done in the usual manner by exact tuple match. Linking will be scored over the same ERE-annotated sub-corpus as described below.

7.2   Corpus-level scoring

Cross-document event coreference (LINK-CORPUS) will be measured using a query /assessment paradigm. NIST will run software that queries a submission and returns a set of responses. LDC will assess the responses.

 A query entry point will specify an (Event Type, Role, Argument, PJ) and request a set of references to the event specified by the entry point. An example entry point (from Figure 1) would be: *Find all references to the event referenced by (CONTACT.MEET, ENTITY,"Mehment Simsek", "… Mehmet Simsek, in Brussels for today's opening of a new accession chapter with the EU…").* Query entry points will always be named or a fully specified temporal expression (e.g. *John Smith, 10-07-2012* and not *the seven victims*) and as much as possible will reference events with clear boundaries (e.g. a meeting and not an ongoing distributed worker's strike).[4] Because systems may choose different spans of text as the canonical argument string for an argument (e.g. John Smith in sentence 1 or John Smith in sentence 2), the software that queries a system's submission will use multiple name string entry points for the same query[5],[6].

The evaluation software will align a query entry point with (a) an event argument assertionand (b) the corpus-level event hopper(s) that assertion is linked to in the corpus. The software will return to the LDC assessors the list of references to the event (grouped by document) that results from taking the predicate justifications for each argument of each document-level event hopper in each returned corpus-level event hopper. The assessors will assess each document-level grouping as

- C: Correct reference to query event *(C will be used if the set of justifications includes **any** mention of query event hopper, even if extraneous information is also included)*.
- W: Wrong
  - ET_Match: A reference to an event of the query event type, but not the specific event (useful for analysis of near misses)
  - No_Match: No reference to an event of the query event type

Performance on the corpus level sub-score will be a function of the system's true positives, false positives, and false negatives compared to a pool of all correct document-level predictions from all

---

[4] Both of these constraints are designed to simplify the task in 2016 but need not be true in the future.

[5] The initial plan is to use RichERE entity annotation to expand the query entry points automatically. Such expansions would include all names in the document. An alternative would be to make use of the notion of canonicalness and use semi-automatic selection of identical strings.

[6] In the future (e.g. 2017) as the event task continue to move to the ColdStart KB formulation, it may be necessary to add a requirement that arguments of events are referenced via global ID (rather than canonical string). This requirement is not being included in 2016 to reduce the burden on participants moving to a corpus-level task.

systems[7]

The LDC assessors will assess the returned documents in random order and will cease assessment if precision drops below 10% (this threshold is subject to revision by the LDC).

7.3 Scoring

A package to automatically validate system output and score is available here: https://github.com/BBN-E/tac-kbp-eal . Systems which are written in JVM-based languages are encouraged to use the classes here directly for representing and writing their output.

7.3.1 Official Metric Details:

The official metric combines three subscores. A, L, and C.

7.3.1.1.1 Subscore1: ARG (unnormalized)

For the event argument extraction sub-score for a document d we use the linear function $\text{ARG}(d) = TP_{EAE}(d) - \beta FP_{EAE}(d)$ for some parameter β, where TRFR true and false positives are defined as above. Intuitively, this corresponds to a model where the user of an EAEL system derives utility 1 from a TP_{EAE} and loses utility β from an FP_{EAE}. Note that this matches the score used in 2015 but 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).

7.3.1.1.2 Subscore2: DOC-LINK (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[8]:

1. Let $S(d)$ be the system-provided TRFR linking for a document d. Let $R(d)$ be the reference TRFR linking derived from ERE, where the ith event frame is a set of TRFRs denoted $R_i(d)$. Define $\widehat{S(d)}$ to be $S(d)$ with all TRFRs not found in $R(d)$ removed (that is, $S(d)$ without EAE false positives).
2. Define $v_Y(x)$ 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:

---

[7] The $S_{\text{link-corpus}}$ sub-score only measures the coherence/completeness at the document of a system's cross-document event coreference decisions. To understand the system's overall KB accuracy, it is necessary to combine the corpus score with the other two sub-scores. However, because the query entry-points are argument driven, it is unlikely that a system could perform well on the cross-document metric without reasonable recall.

[8] 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.

16

    a. If x is not in Z, $f(x) = 0$

    b. If $x \in Z$ and $v_Y(x)$ and $v_Z(x)$ are empty, then $f(x) = 1$.

    c. Otherwise, let $p_{Y,Z}(x)$, the precision, be $\frac{|v_Z(x) \cap v_Y(x)|}{|v_Z(x)|}$. Let $r_{Y,Z}(x)$, the recall, be

    $\frac{|v_Z(x) \cap v_Y(x)|}{|v_Y(x)|}$. $f_{Y,Z}(x) = \frac{2 p_{Y,Z}(x) r_{Y,Z}(x)}{p_{Y,Z}(x) + r_{Y,Z}(x)}$

4. Let $U_X(d)$ be the union of all event frames in X. We define DOC-LINK(d, R, L) as

    $\sum_{x \in U_R(d)} f_{\widehat{S,R}}(x)$. Intuitively, it is the sum of the link F scores for each TRFR present in the gold standard.

7.3.1.1.3 Subscore3: CORPUS-LINK (C)

Each corpus-level query will be scored separately as $\max(0, \frac{TP_{query} - \gamma FP_{query}}{TP_{query} + FN_{query}})$ where TP$_{query}$ is the number of documents returned by the system for the query which were judged correct, FP$_{query}$ is the number of documents returned by the system for the query which were judged incorrect, and FN$_{query}$ is the number of documents returned by other systems which were assessed as correct but not returned by this system.

The final CORPUS-LINK score is the mean over the score for all queries.

7.3.1.1.4 Aggregating Scores

Define the scores over an ERE-annotated corpus D as ERE$(D) = \frac{1}{N_{TRFR}} \sum_{d \in D} [\lambda_0 max(0, \text{ARG}(d)) + \lambda_1 \text{DOC-LINK}(d)]_0$ where N$_{TRFR}$ is the number of TRFRs in the gold standard for D. Note that while ARG can be negative, we clip it to 0 on a per-document basis.

Define ERE$'$ as the median ERE over 1000 bootstrap samples over the ERE annotated evaluation corpus. Define CORPUS-LINK$'$ as the median value of the corpus linking score over 1000 bootstrap samples over the evaluation queries. The final score is RANK2016 $=$ ERE$' + \lambda_2$CORPUS-LINK$'$

7.3.1.1.5 Official Ranking Score

For the official ranking score, we will use $\beta = \frac{1}{4}, \lambda_{0,1,2} = \frac{1}{3}, \gamma = \frac{1}{4}$ to weigh all components roughly equally[9] and to encourage high recall while maintaining reasonable precision. 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. A good value for γ is particularly uncertain. 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 RANK-2016 as described above We will report for each rank the fraction of the cross-product of ERE corpus samples and query set samples on which it outperforms each other rank.

7.3.1.2 *Additional Diagnostic Metrics*

---

[9]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.

TAC 2016 Event Argument and Linking Task Description: **DRAFT**
This submission does not contain any material which may be restricted under EAR/ITAR

As in 2015, we will provide several diagnostic metrics, for example:

- Each sub-score independently
- Each sub-score independently by language
- The overall score by language
- ARG and an F1-based version of ARG ignoring the impact of realis

7.4   Training 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, but including: LDC2015E78, LDC2015E105, LDC2015E112, LDC2015E68, LDC2015E29, LDC2015E107)
- Event Nugget and Event Nugget Coreference training data

Participants will also be provided with the assessments from the 2014 and 2015 Event Argument Task.

7.5   Evaluation Period Resources for Participants

While cross-document entity coreference is not a strict requirement of this task, cross-document coreference is expected to aid in event coreference. The organizers hope to provide system EDL & ColdStart output for the full 90K document corpus from those systems who have submitted submissions to either the ColdStart or T-EDL evaluation windows.   Output will be provided with system ID to allow participants to use historical information to decide which systems to trust. Participants are also encouraged to make use of the resources listed in http://nlp.cs.rpi.edu/kbp/2015/tools.html for cross-document entity coreference.

7.6   Submissions and Schedule

7.6.1   Submission

Systems will have up to one week to process the evaluation documents.[10] 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.

7.6.2   Schedule

February 2015	Task definition released. Pre-existing resources (ACE data, ERE data, 2014 EA data) available to participants as they sign up

---

[10] Participants with limited computing resources are reminded that many cloud computing services provide sufficient computing resources to run the evaluation in their "free" tiers.

March	Scoring software release
May	Dry Run (optional)
June 1, 2015	Final versions of guidelines and software released
August 2015	Evaluation Period

7.7   References

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

ACE Task Guidelines: https://www.ldc.upenn.edu/collaborations/past-projects/ace
http://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

RichERE. Rich ERE guidelines are distributed by LDC in the documentation of RichERE annotation releases (e.g. LDC2015E29).