Overview of Linguistic Resources for the TAC KBP 2015 Evaluations: Methodologies and Results

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
{joellis, jgetman, foredana, neilkus, zhiyi, bies, strassel}@ldc.upenn.edu

Abstract

Knowledge Base Population (KBP) is an evaluation track of the Text Analysis Conference (TAC), a workshop series organized by the National Institute of Standards and Technology (NIST). In 2015, TAC KBP's seventh year of operation, the evaluations focused on four tracks targeting information extraction and question answering technologies: Tri-lingual Entity Discovery & Linking, Cold Start, Event Argument Linking, and Event Nuggets. Linguistic Data Consortium (LDC) at the University of Pennsylvania has supported TAC KBP since 2009 and continued again in 2015, developing, maintaining, and distributing new and existing linguistic resources for the evaluation series, including queries, human-generated responses, assessments, and tools and specifications. This paper describes LDC's resource creation efforts and their results in support of TAC KBP 2015.

1 Introduction

TAC KBP (McNamee et al. 2010), started in 2009, focusing on information extraction and question answering technologies and evolving primarily from two other programs - TREC Question Answering (Dang et al. 2006) and Automated Content Extraction (ACE) (Doddington et al. 2004).

Since 2009 Linguistic Data Consortium (LDC) has been the primary data provider for the evaluation series, developing and distributing training and evaluation datasets as well as tools and specifications. In 2015, LDC created a total of 56 new data sets in support of the KBP evaluations, 4 more than were produced for the 2014 evaluations. Eleven of these releases were developed early in 2015, well before the start of the evaluations. In an effort to ease the distribution and use of existing KBP data, LDC condensed 74 TAC KBP data sets produced from 2009 – 2014 into just 11 corpora. The remaining 45 corpora created this year include training and evaluation data releases for participants, preliminary releases to coordinators for data previews, updates to existing releases to improve quality and add new data, and supplemental releases distributed only to coordinators to support their own data production efforts.

TAC KBP 2015 comprises 4 separate evaluation tracks – Tri-Lingual Entity Discovery & Linking, Cold Start, Event Argument Linking, and Event Nugget. This paper describes the data creation efforts for those evaluations. Sections 2 through 5 discuss the procedures and methodologies for data selection, query development, annotation, and assessment for all of the 2015 tracks. Section 6 offers concluding remarks. The appendix lists all final datasets released by LDC in support of TAC KBP 2015.

2 Tri-Lingual Entity Discovery & Linking

For 2015, the Entity Discovery & Linking task from 2014 was broadened from mono-lingual to tri-lingual coverage, expanding from English to include Spanish and Chinese. The goal of the Tri-lingual Entity Discovery & Linking task (TED&L) is full entity extraction from a collection of documents in the three languages, followed by linking all entities to the Knowledge Base (KB) and clustering all NIL entities. Other changes from the 2014 monolingual task include the addition of location (LOC) and facility (FAC) entity types as well as nominal mention types (though the latter addition is incorporated for English only).

2.1 Knowledge Base

After much deliberation with previous KBP participants, coordinators, and sponsors, BaseKB (http://basekb.com/) was selected as the new Knowledge Base early in the 2015 planning cycle. BaseKB is a curated subset of Freebase converted to RDF and reduced to about half the number of facts though still containing over a billion facts about more than 40 million subjects. As a triple store, BaseKB also allows systems to interact with the KB as a graph. Additionally,

Since BaseKB is a triple store, the first step required to support TED&L annotation was to create a human-readable version, with "pages" that annotators could review and link to as necessary. Producing a human-readable version of the KB proved challenging for several reasons, including

unintelligible values for subjects, predicates, and objects and complex objects which required unpacking in order to find actors for simple relations (e.g. 'marriage events' that indicated spouses). Despite these and other difficulties, a version of the human-readable KB was eventually produced and the algorithm by which it was created was distributed to participants.

2.2 Source Data

The 15 documents used in the pilot TED&L task were selected by the track coordinator specifically for their ability to exercise cross-lingual, cross-document entity clustering. The new documents selected for the training and evaluation sets, however, were required to meet a stricter set of requirements. In addition to containing cross-lingual, cross-document entity mentions, source documents for the training and evaluation data sets were required to be taken from multiple, various, and relatively recent sources in order to test systems' abilities at handling a variety of content and entities mentioned in more recent news. This required collection of a new data set for TED&L.

Both the TED&L 2015 training data source corpus (444 documents) and evaluation data source corpus (499 documents) are comprised of approximately half newswire documents and half discussion forum threads in all three of the target languages. LDC annotators scouted URLs for sources pertaining to a set of approximately 10 topics in recent international news. Given the requirement for the corpus to include a small number of documents from each of many different sources, the data was harvested and processed using a custom approach rather than relying on LDC's existing web collection framework, which is optimized for much larger data volumes. Managing intellectual property rights issues for hundreds of separate data sources also required a customized approach, and source data was distributed in an intermediate data format (LTF), along with tools to produce the expected KBP XML.

2.3 Gold Standard Data Production

The TED&L gold standard training and evaluation data was produced by native speakers of the three target languages who exhaustively identified and classified all valid, named mentions of facilities (FAC), geopolitical entities (GPE) , locations (LOC), organizations (ORG) and persons (PER) occurring within the source corpus. For the English-version of the task, the heads of singular nominal mentions of person entities (victim, suspect, wife, etc.) were also annotated, as were titles in order to help systems distinguish between the two. The guidelines directed annotators to use context and to tag for meaning in order to confidently identify the intended referent of any entity mention within the text.

Within-document coreference was performed concurrently with entity discovery annotation and, for each cluster of entity mentions, annotators also indicated entity type. Each entity cluster was then linked to a node in the KB or marked it as NIL (to indicate that the entity did not have a node in the KB) or Unknown (to indicate that insufficient information about the entity was available in the source document to confidently identify whether or not it had a node in the KB). Annotators were also prompted to indicate whether or not an online search of the web and/or extensive reading of the prose description within a KB page were necessary in order to determine the identity of an entity cluster.

Following the completion of document level annotation, a final, cross-document, cross-language coreference of all NIL entities was performed by an experienced English annotator. Using English descriptions about the referents for each non-English cluster, the annotator used a combination of sorting and informed searching in order to identify mention clusters from different documents that needed to be collapsed into a single cluster.

As discussed earlier, the format and complexity of BaseKB made it difficult to prepare a human-readable version of the KB for annotators to use. However, the other, more significant, challenge faced was indexing the human-readable version so that it could be searched by annotators. Despite multiple attempts at indexing, none supported productive searching by annotators. As an example, in the second version of the index we created, when an annotator searched for "united states" in the KB, the actual page for the US was approximately the 650th result. Since the timeline of the evaluations did not allow for extended experimentation, LDC eventually developed a workaround in which annotators searched live Freebase (freebase.com) for linking purposes instead of the human-readable derivative of BaseKB. Since Freebase is a superset of BaseKB, any entity IDs found by annotators that did not also exist in Base KB were replaced with NIL IDs during the production of output.

	Training Data	Evaluation Data
Total mentions	30,838	32,533
ENG	13,545	15,645
CMN	13,116	11,066
SPA	4,177	5,822
Total equivalence classes	5,744	7,235
ENG	2,702	3,190
CMN	1,827	2,139
SPA	739	1,363
ENG/CMN	170	159
ENG/SPA	96	123
CMN/SPA	38	38
ENG/CMN/SPA	172	223

Table 1: TED&L 2015 Data Volumes

2.4 Future Efforts

While the approach described above was sufficient to allow us to support the 2015 TED&L evaluation, improvements to the process are required if KBP continues to use BaseKB in future. Primarily, since there are no guarantees that Freebase will remain freely accessible online, a new approach to developing a human-readable version of BaseKB that can be reliably searched by annotators is needed. LDC and NIST have recently begun discussing a second attempt at producing a version of BaseKB that can be searched and viewed by annotators for use in future evaluations.

3 Cold Start

From a participant perspective, the core Cold Start task may not have appeared to have changed much from 2014 to 2015. Indeed, this year’s task description even indicated such, describing the 2015 KB variant as “the same as the 2014 Cold Start Knowledge Base task”. However, while seemingly subtle, the changes in query requirements for Cold Start in 2015 required drastic changes to the process by which Cold Start queries were developed this year as compared to 2014.

3.1 Source Document Pools

For 2015, Cold Start returned to the requirement that all the source documents remain unexposed until the start of the evaluation. In 2014, LDC had acquired two pools of unreleased source documents (one being a collection of approximately 57,000 New York Times documents from 2013 and the other consisting of over 1 million discussion forum threads collected from multiple sources in 2014). Since these collections of unreleased documents were already available for use in KBP, annotators started developing the Cold Start queries and manual run from them as soon as the task specifications and the annotation UI were in place. However, in order to add diversity to the collection and to provide annotators with a larger set of newswire documents with which to develop the queries and responses, LDC acquired another collection of newswire documents from the Xinhua news agency, totaling over 86,000 documents, covering the same time period. Negotiations for the Xinhua documents were protracted, however, and the documents were not acquired until late in the query development process. As such, the additional documents were added to the set of sources later in the process of query development, at which time annotators focused on generating queries and responses from them in order to diversify the set of source inputs.

3.2 Cold Start Query Development and Manual Run

On the surface, Cold Start queries were the same in 2015 as they had been in previous evaluations, consisting of an initial ‘entry point’ entity and one or two TAC KBP slots by which the entity would be connected to others in the corpus. For example, given the text “Jane Doe is the president of the School of Arts and Sciences at the University of Pennsylvania”, one could produce the following query (but with the potential for more responses at each level depending on whether or not Jane Doe was indicated in the corpus as having been an employee at other organizations):

“Jane Doe”
 per:employee_of
  “School of Arts and Sciences”
  org:parents
   “University of Pennsylvania”

From 2012 – 2014, there were no requirements for interrelatedness between Cold Start queries; while it had been possible and acceptable for queries to share entry point entities, there was no requirement for such, which allowed for queries to be created independently of one another (with the exception of ensuring no queries were duplicated). However, for 2015, Cold Start queries were required to have a high degree of overlap with respect to entry point entities and, whenever possible, references to the entities were to derive from separate source documents. Additionally, ambiguous references to query entities, such as aliases or underspecified mentions, were to be used as entry points whenever possible. These updates to query requirements were made primarily to better align Cold Start with the Slot Filling task, since the latter was being entirely subsumed by the former for the 2015 evaluations, and to ensure adequate challenge for Cold Start - Entity Discovery, a new variation for 2015, which will be discussed later.

To meet the challenges presented by the new query requirements, LDC created a new Cold Start query development UI for 2015 data creation efforts. The new interface allowed annotators to more easily develop the necessary webs of interrelated queries by allowing for the concurrent development of multiple queries that shared an entry point entity As was done in previous years, the new Cold Start query development process started with searches focusing on key words related to the KBP slots in order to find entities connected by the defined relations. For example, annotators might have searched for “hired” or “resigned” to develop queries that would generate fillers for the *per:employee_or_member_of* slot. However, once an initial ‘seed’ annotation such as the above was found, query developers searched for other mentions of the connected entities in the corpus to determine which would be used as the query’s entry point entity. Since most Cold Start slots can be inverted, either entity in the seed annotation could be used as the query entity and so annotators researched both on the web and in the three pools of source documents in order to determine which would be most productive, which had more references throughout the corpus, and which was referred to by ambiguous name strings. Note, however, that less productive query entities were also selected if they offered opportunities to balance query entity types (PER, GPE or ORG), slot types (name, value or string), and document genres (newswire or MPDF). Once a query entity was selected, annotators captured up to five named mentions of it from different source documents, opting for ambiguous mentions whenever they were available. The final step for annotators in producing a set of queries from an entry point entity was to select slots that produced valid responses. To do so, query developers first selected slots for hop-0 by searching the corpus for all slots which, when paired with the entry-point entity, would produce at least one valid filler. Having done so, all entity fillers produced from the hop-0 slots were also investigated and paired with another KBP slot when doing so produced valid hop-1 fillers.

After annotators had completed development of queries and responses for Cold Start, the next step in the pipeline was satisfying another new requirement for 2015 - the development of null queries, those without known answers in the corpus. In order to save time and annotator effort, null queries were generated automatically by copying the productive, annotator-produced queries and then replacing the selected slots with alternates of the same filler and entity type. For two-hop queries, only slot-1 was replaced. It is important to note that, while null queries were intended to be queries without valid responses, the process for producing them, which was developed by LDC in collaboration with Cold Start coordinators, was not guaranteed to produce queries without responses; it was only a means by which to rapidly produce new queries that were not *guaranteed* to produce valid responses (as they had not been reviewed by any humans). After null queries were produced, they and the set of productive queries were mixed together and randomized before being assigned query IDs.

Total queries	2,557
Total productive (i.e. non-null) queries	1,327
Total unique entry-point entity mentions	1,148
Total manual run responses	2,218
Hop-0 responses	1,460
Hop-1 responses	758

Table 2: Cold Start 2015 Data Volumes

3.3 Source Corpus Selection

The last step in creating the set of inputs for the Cold Start evaluation was selecting the final set of documents included in the source corpus. After development of the Cold Start queries and manual run was complete, the source documents from which all of the annotations were drawn were first added to the list of documents to be used in the evaluation. Afterward, additional documents from the three original source document pools were added until the target of approximately 50,000 sources was reached. The additional documents were selected using fuzzy string matches against all of the entities included in the queries and manual run, with some effort taken to balance the representation of the name strings in order to avoid over-representation of a few names in the documents. Excessively long documents were also avoided.

3.4 Entity Discovery

In order to support the Entity Discovery task (ED), a new variant for Cold Start in 2015 that sought to focus systems on the challenges of identifying and coreferencing valid entity mentions in text, annotators were required to find and exhaustively select and coreference all named mentions of the three valid entity types appearing in the ED source corpus. LDC first developed a sample ED data set (LDC2015E72) from two source documents in the 2014 Cold Start evaluation corpus. As the primary purpose of the sample release was to provide coordinators with a means to develop scoring software, the two documents – one newswire and one discussion forum thread – were specifically selected to exercise cross-document coreference challenges.

Source documents for the ED evaluation were required to be a subset of those utilized by the Cold Start evaluation queries and manual run. Additionally, like the sample data, the evaluation source corpus had to include entities that were mentioned in multiple documents but with the added challenge that at least some of the mentions had to be ambiguous. To meet this requirement, at the start of data development for Cold Start queries and the manual run, annotators were instructed to focus heavily on generating queries with entry point entities who were mentioned in multiple documents and who were referred to ambiguously in some of those documents. This process was monitored closely and, soon after 200 documents had been annotated for the Cold Start queries and manual run, those same documents were selected for use as the Entity Discovery evaluation source corpus, following additional review to balance genres.

Total entity mentions	7,718
Total PER mentions	3,335
Total ORG mentions	2,005
Total GPE mentions	2,378

Table 3: Cold Start-Entity Discovery Data Volumes

3.5 Cold Start Assessment

For the first time since the start of Slot Filling in 2009 and Cold Start in 2012, systems returned more responses in the 2015 Cold Start evaluation than could be assessed in time for reporting scores in advance of the workshop. As a result, coordinators needed to develop a process for down-selecting pools of responses that would be included in assessment. To support the development of this process, LDC assessed multiple batches of responses to allow for coordinators to analyze results more quickly and determine whether any changes in the selection process were needed. To date, LDC has assessed 10 batches of responses, each consisting of both a hop-0 and hop-1 pool, for a combined total of over 30,000 assessments.

Before assessment began, annotator training and testing was performed as a preliminary step. After an initial training session and guidelines review, candidate assessors were required to complete an assessment screening kit, which contained 50 sample responses selected from past KBP evaluations. Assessors were required to assess every slot in the test kit and achieve 90% or higher accuracy for all slots. Those who passed the test went on to assess and coreference responses.

The actual task of assessment varied only slightly in 2015 from what had been conducted in the previous year. Fillers were marked as correct if they were found to be in-line with the slot descriptions and supported in the provided justification string(s) and/or its surrounding content. Fillers were assessed either as wrong if they did not meet both of the conditions for correctness or inexact if insufficient or extraneous text had been selected for an otherwise correct response. Justification was assessed as correct if it succinctly and completely supported the relation, wrong if it did not support the relation at all (or if the corresponding filler was marked wrong), inexact-short if part but not all of the information necessary to support the relation was provided, or inexact-long if it contained all information necessary to support the relation but also a great deal of extraneous text. Changes for the 2015 version of the task included the removal of the 'Ignore' assessment as responses with justification comprised of more than 600 characters in total were simply never passed on to assessors to review. Additionally, responses marked as 'Inexact' were coreferenced with 'Correct' responses whereas in previous years only correct responses were included. As with the development of the manual runs, after first passes of assessment were completed, quality control was performed on the data by annotators who reviewed the work of their peers and flagged potentially problematic assessments for additional review.

	Total	Newswire	MPDF
Responses	30,654	15,948	14,706
Correct	26.70%	29.70%	23.50%
Wrong	68.80%	65.20%	72.80%
Inexact	4.50%	5.10%	3.70%

Table 4: Cold Start Assessment Results

3.6 Results

Scores for LDC's manual run went down in 2015, from 91% precision and 46% recall (62% F1) in 2014 to 85% precision and 32% recall (46% F1) in 2015.

Track	Precision	Recall	F1
2014 Cold Start	91%	46%	62%
2015 Cold Start	81%	19%	30%

Table 5: LDC's Scores for Slot Filling and Cold Start

We believe that the primary causes of the drop were the challenges presented by the new query requirements in 2015. While the new Cold Start query development approach allowed for easier tracking of queries that shared a common entry point entity, it also complicated the process of developing the manual run. Since the new approach required annotators to develop multiple queries concurrently, we believe that this caused for less focus to be given to fleshing out responses for the manual run.

That said, however, the number of queries produced for the 2015 evaluation as well as the timeline in which they were produced were certainly factors as well. For the 2015 Cold Start evaluation, LDC annotators created 1,327 queries (not considering hop-1 portions of the queries separately). By comparison, we developed a total of 750 queries for the Cold Start, Slot Filling, and Sentiment Slot Filling evaluations combined in 2014. Should Cold Start data requirements for 2016 be similar to those for this year, a second step to the query development/annotation task in which another group of annotators conduct a second, time-limited pass for the manual run to ensure better coverage of query responses could improve results.

4 Event Argument Linking

Event Argument Linking (EAL) is a further developed version of the Event Argument Extraction task first run in 2014. In EAL, annotators and systems extracted mentions of entities from unstructured text and indicated the role they played in an event as supported by text. Critically, the extracted information had to be suitable as input to a knowledge base and so annotators and systems produced tuples indicating the event type, the role played by the entity in the event, and the most canonical mention of the entity itself from the source document. Event argument tuples were then "linked" with other tuples into Event Hoppers, indicating that the tuples played a role in the same event or events. Event Argument Linking was made up of three separate processes – source document selection, manual run development, and assessment.

4.1 Document Selection

Documents served as queries in EAL and so the first step for annotators in developing data for the task was to perform targeted searches over two sets of previously unreleased documents (one set of newswire documents and one set of discussion forum threads). Documents were selected from the same pool from which the 2014 EAE source corpus was selected, though steps were taken to ensure annotators did not select documents used and released in 2014. A total of 250 NW and 250 MPDF documents were selected by annotators.

Documents were selected based on the criteria that they contained at least one

"Actual" mention of one of the specified event types along with valid arguments for the event. "Actual" events, as defined in EAL, include those that happened in the past or those that are ongoing in the present. Documents with a variety of event types were primarily sought after, though documents providing mentions of generally less common event types were also selected for inclusion.

Upon finding a promising document, selectors reviewed the text closely and tallied the number of unique event mentions of each event type that was included. Such tallies helped ensure that all of the targeted event types were at least reasonably well-represented in the corpus of documents selected for the EAL evaluation. Specifically, annotators attempted to select documents such that a minimum of 10 Actual event instances of every event type within each genre (i.e. 10 in newswire and 10 in discussion forum) would occur within the corpus overall.

While performing document reviews, annotators also searched for certain undesirable qualities that would prevent a document from being included in the corpus. Most notably, discussion forum documents with more than a small amount of newswire quotation were avoided with the aim of selecting discussion forum data actually comprised of informal content.

4.2 Manual Run

Following the evolution of EAE into EAL, a new tool was developed in 2015 for LDC's manual run, one which improved and streamlined the approach to the event argument extraction portion of the task (compared with annotation approaches used in 2014) and which would handle the addition of the argument linking portion of the task, new in 2015.

LDC performed the manual run over a subset of the 500 document EAL source corpus. Using the event tallies produced during document selection, LDC further downselected the source corpus to the 300 documents that produced the largest event tally possible while also balancing the mix of event types as much as possible. Priority was given to keeping the event types mixed, and we attempted to ensure that each event type was still represented at least 10 times/genre across the corpus overall within the 300 document sub-corpus.

For each of the 300 documents in the EAL evaluation over which the LDC manual run was performed, an annotator had a maximum of sixty minutes to annotate all valid, unique event arguments within that document and to decide with which event hoppers to link each and every annotated event argument. Following the initial round of annotation, a quality control pass was conducted over the manual run data to flag any event arguments or linking decisions that did not have adequate justification in the source document, or that might be at variance with the current guidelines. These flagged annotations were then adjudicated by senior annotators.

4.3 Assessment

For the assessment of EAL responses produced during the evaluation, LDC used an online tool developed and graciously provided by BBN. EAL assessment was comprised of three subtasks: entity coreference, response assessment, and argument linking.

After initial training, candidate assessors were required to complete three assessment training kits and their responses were then compared to a set of gold standard versions of the kits completed by a senior annotator. Each assessor then received further, individual training to focus on the areas in his or her training kits that were at odds with the gold standards.

Once assessors had completed their training, they began work on production kits. The first step in each EAL assessment kit was to perform entity coreference on all responses returned by systems and LDC for a given document. This included correct responses, inexact responses and wrong responses.

Following the completion of entity coreference, assessors moved on to response assessment. Each response generated for EAL received six judgments by an assessor. Event type, argument role (the role that a response played in its matched event), base filler (the mention of the argument included in the justification) and canonical argument string (the 'most complete' mention of the argument from the document) were all marked as 'correct' if they were found to be supported in the sources and in-line with the definition of the relevant event and argument role. Responses were considered 'wrong' if they did not meet both of the conditions for correctness and 'inexact' if overly insufficient justification was provided or extraneous text was selected for an otherwise correct response. Additionally, each response was given a 'realis' judgment, by which a general judgment regarding the modality of the event argument was made ('Actual' if the event clearly occurred in the past or present, 'Generic' if the event was generic in nature – e.g. "I go to the store on Sundays", and 'Other' if the event could not neatly be described as one of the other two categories). Lastly, assessors also marked the canonical argument strings as either 'name' or 'nominal' to indicate the type of mention.

As assessment was completed, quality control was performed on the data. Senior annotators reviewed the work of assessors and made corrections to assessment kits and, for each correction that was made, the reviewer followed up with the original assessor to clarify the correction. For certain classes of potential errors, BBN produced QC reports for senior annotators to review while performing quality control. Following the completion of quality control for a given document, the senior annotator who had performed the QC for that document then performed the document's argument linking step as well, which was comprised of deciding how correct and inexact responses should be grouped together in event hoppers.

Number of arguments returned by event type	Number of Event Types in range	Percent of Manual Run
>300	3	20%
200-299	8	39%
100-199	8	23%
<99	13	18%

Table 5: Event Argument Linking Data Volumes

4.4 Results

LDC's precision remained consistent with last year's effort but recall was again lower this year in the Event Argument track than in other past and current KBP tracks, though preliminary results show that recall and overall score were quite improved.

Track	**Precision**	**Recall**	**F1**
2015 EAL (preliminary)	76%	40%	52%
2014 EAE	76%	28%	41%

Table 6: LDC's Event Argument Extraction Scores

Primarily, we believe this to be caused by the relatively small amount of time during which annotators have to find, annotate and link all event arguments in a document. In Rich ERE, a comparable event annotation task, annotators generally spend at least 3.5 hours performing event annotation on a single document of a size comparable to the documents in the EAL source corpus.

5 Event Nugget

The Event Nugget (EN) evaluation in 2015 sought to evaluate system performance in detecting and coreferencing references to events in text. An event 'nugget', as defined by the task, includes a text extent, a classification of event type and subtypes, and an indication of whether realis mood was used to describe the event.

EN started as a pilot evaluation within the DEFT program in 2014. However, for the current version of the task conducted as part of the 2015 TAC KBP evaluations, many updates were made, including a redefinition of valid text extents to align with those used in the Rich Entities, Relations, and Events (ERE) data (Song, et al., 2015) as well as the addition of event coreference.

5.1 Changes from the 2014 Event Nugget Pilot Evaluation

In 2014, LDC supported a pilot run of the Event Nugget (EN) evaluation by actively participating in task definition discussions and, subsequently, developing both training and evaluation data. This preliminary version of the EN task adapted the event annotation guidelines from LDC's Light ERE annotation task (a simplified version of Rich ERE that does not include event coreference) by incorporating modifications by coordinators at CMU that focused on the text extents establishing valid references to events, clarifications on transaction event types, and the additional annotation of event attributes.

The 2014 EN pilot sought to serve two purposes, one was to measure event detection systems of some performers in the DEFT program. The other purpose of the pilot, however, was to test run the evaluation framework before opening it up to the full TAC KBP community. As the pilot was considered a success, EN was added to the roster of evaluation tracks for TAC KBP 2015 but with some modifications to incorporate lessons learned from the pilot and to better align with both the other TAC KBP 2015 event-related evaluation – Event Argument Linking – as well as the Rich ERE data provided to KBP participants for training purposes.

For the 2015 EN evaluation, event 'triggers' – the textual extent indicating a reference to a valid event – was redefined as the smallest, contiguous extent of text (usually a word or phrase) that most saliently expresses the occurrence of an event. Additionally, annotators for the 2015 data were allowed to 'double tag' event triggers in order to indicate that a given text extent referred to more than one event and was usually used to indicate the presence of an inferred event. For example, given the following text:

> *Cipriani was sentenced to life in prison for the murder of Renault chief George Besse in 1986 and the head of government arms sales Rene Audran a year earlier.*

the word "murder" would be the trigger for two *Life-Die* events, one with the victim "George Besse" and the other with "Rene Audran" as well as two *Conflict-Attack* events, one occurring in 1986 and the other in 1987.

This year's evaluation also added a new event type (Manufacture) and four new subtypes – *Movement.TransportArtifact, Contact.Broadcast, Contact.Contact, Transaction.Transaction* – which aligned the EN event ontology with that of Rich ERE. Importantly, the EN annotation task also adopted a new approach for applying Contact event subtype categorizations, which had been developed for Rich ERE data creation efforts. Instead of having annotators categorize the subtypes directly, Contact event mentions were labeled with attributes to describe formality (Formal, Informal, Can't Tell), scheduling (Planned, Spontaneous, Can't Tell), medium (In-person, Not-in-person, Can't Tell), and audience (Two-way, One-way, Can't Tell). Contact event subtypes were automatically generated based on the applied attributes.

The final change to the EN evaluation as compared to the 2014 pilot was the added requirement of event coreference. Again taking from the Rich ERE task, EN addressed the challenge by adopting the notion of 'event hoppers', a more inclusive, less strict notion of event coreference as compared to previous approaches. Following this approach, event mentions are added to an event hopper when they "feel" coreferential to an annotator, even if they do not meet a strict event identity requirement. Event nuggets could be placed into the same event hoppers even if they differed in temporal or trigger granularity, their arguments were non-coreferential or conflicting, or even if their realis attribute differed.

5.2 Event Nugget Annotation

Given the level of changes for the 2015 EN evaluation as compared to the pilot task, preliminary efforts for support involved re-annotating the pilot evaluation data to reflect the new requirements, a joint effort taken up by CMU and LDC. Fifty documents from the pilot data were reviewed and updated by CMU annotators and then given an additional quality control pass by LDC, using a slightly modified version of the Rich ERE annotation tool. The remaining training files were annotated solely by LDC. In addition to updating the 2014 pilot EN data, new and existing Rich ERE data was made available to participants for the 2015 EN evaluation.

Source data for the 2015 Event Nugget evaluation was a subset of the documents selected for Event Argument Linking (EAL), which had been manually selected to ensure coverage of all event types and subtypes for that evaluation. Although 300 documents were manually annotated for EAL, only 200 were used in the EN evaluation, as requested by track coordinators. The set of documents was down-selected based primarily on token count (with shorter documents preferred) and then by balancing the two document genres and representation of all of the event types (at least 5 of each event type was included in the final set). Tokenization of the source documents was also provided but, unlike the 2014 data, in which annotation was performed on pre-tokenized text, the processed was performed as a post-annotation procedure, using tool kits provided by evaluation coordinators.

In order to reduce the impact of low recall on annotation consistency, which had proven problematic in the pilot and in previous event annotation efforts (Mitamura, et al., 2015), gold standard EN data was conducted by first having two annotators perform event nugget annotation (which included the creation of event hoppers) independently for each document followed by an adjudication pass conducted by a senior annotator to resolve disagreements. The EN annotation team consisted of nine annotators, six of whom were also adjudicators and care was taken to ensure that annotators did not adjudicate their own first pass files. Following adjudication of all documents, a corpus-wide quality control pass was also performed. Since the data was created in the existing ERE tool, track coordinators provided tools to convert the ERE output format to the different files needed to support the evaluations.

	Genre	Files	Nuggets	Hoppers
Totals	NW/DF	360	12,976	7,460
Training	NW	81	2,219	1,461
Training	DF	77	4,319	1,874
Eval	NW	98	3788	2440
Eval	DF	104	2650	1685

Table 7: Event Nugget Data Volumes

5.3 Results

Subsequent analysis of inter-annotator agreement in the EN data indicates that several challenges remain to be addressed but annotation consistency is generally in line with what we expect due to the complex nature of event recognition. Although the changes in the approach to the annotation task that were described above appear to have made some improvements, they were not significant. Regarding event coreference, which was new in 2015, annotator consistency was 67.63%. Figure 2 compares the overall inter-annotator agreement[1] on first pass Event Nugget annotation in 2015 and 2014.

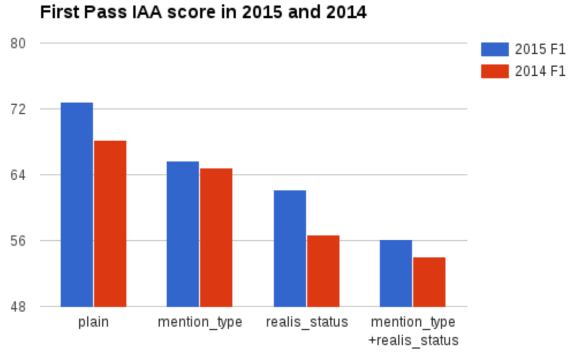

Figure 2. First pass EN annotation inter-annotator agreement

6 Conclusion

This paper discussed the linguistic resources produced in support of the TAC KBP 2015 evaluations, from the planning processes and data creation efforts to descriptions of the datasets and analysis of how results compared to previous efforts. LDC support of TAC KBP in 2015 included contributions to task descriptions, data curation and distribution, source corpus expansion, and creating or revising existing data development procedures to accommodate new or modified evaluations. Future work will include repackaging and updating documentation to make the data created this year more readily useable in the future by system developers who may be unfamiliar with the KBP evaluations. The resources described in this paper will be published in the LDC Catalog, in order to make the corpora available to the wider research community.

7 References

Hoa T. Dang, Jimmy Lin, and Diane Kelly. 2006. Overview of the TREC 2006 Question Answering Track. In *Fifteenth Text Retrieval Conference (TREC 2006) Proceedings*, Gaithersburg, MD.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. Automatic Content Extraction (ACE) program - task definitions and performance measures. In *Proceedings of the Fourth International Language Resources and Evaluation Conference*, Lisbon, Portugal.

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. NYU's English ACE 2005 System Description. In *Proceedings of the ACE 2005 Evaluation/PI Workshop*, Washington D.C., U.S.A.

Paul McNamee, Hoa T. Dang, Heather Simpson, Patrick Schone, and Stephanie

[1] IAA scores were computed using scorer_v1.6.py, the same scorer for the Event Nugget evaluation (Liu, 2015). The scorer is available on https://github.com/hunterhector/EvmEval/zipball/master

M. Strassel. 2010. An Evaluation of Technologies for Knowledge Base Population. In *Proceedings of the Seventh International Language Resources and Evaluation Conference (LREC),* Valleta, Malta.

Heather Simpson, Stephanie Strassel, Robert Parker, and Paul McNamee. 2010. Wikipedia and the Web of Confusable Entities: Experience from Entity Linking Query Creation for TAC 2009 Knowledge Base Population. In *Proceedings of the Seventh International Language Resources and Evaluation Conference (LREC),* Valleta, Malta.

Ramshaw, L., E. Boschee, M. Freedman, J. MacBride, R. Weischedel, A. Zamanian. SERIF Language Processing — Effective Trainable Language Understanding. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation,* editors J. Olive et al., pp.626-631, Springer, 2011.

Zhengzhong Liu, Teruko Mitamura, Eduard Hovy. 2015. Evaluation Algorithms for Event Nugget Detection: A pilot Study. 3*rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation,* at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015).

Teruko Mitamura, Yukari Yamakawa, Sue Holm, Zhiyi Song, Ann Bies, Seth Kulick, Stephanie Strassel. 2015. Event Nugget Annotation: Processes and Issues. *3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation,* at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015).

Zhiyi Song, Ann Bies, Tom Riese, Justin Mott, Jonathan Wright, Seth Kulick, Neville Ryant, Stephanie Strassel, Xiaoyi Ma. 2015. From Light to Rich ERE: Annotation of Entities, Relations, and Events. *3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation,* at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2015).

Appendix A: Data Available to KBP Performers in 2015

Table 1: 2009 – 2014 TAC KBP Data Sets Condensed

Catalog ID	Title	Release Date
LDC2014T16	TAC KBP Reference Knowledge Base	*all pre-2015 data*
LDC2015E17	TAC KBP Chinese Entity Linking Comprehensive Training and Evaluation Data 2011 - 2014	*all pre-2015 data*
LDC2015E18	TAC KBP Spanish Entity Linking - Comprehensive Training and Evaluation Data 2012 - 2014	*all pre-2015 data*
LDC2015E19	TAC KBP English Entity Linking - Comprehensive Training and Evaluation Data 2009 - 2013	*all pre-2015 data*
LDC2015E20	TAC KBP English Entity Discovery and Linking - Comprehensive Training and Evaluation Data 2014	*all pre-2015 data*
LDC2015E22	TAC KBP English Event Argument Extraction - Comprehensive Pilot and Evaluation Data 2014	*all pre-2015 data*
LDC2015E45	TAC KBP Comprehensive English Source Corpora 2009-2014	*all pre-2015 data*
LDC2015E46	TAC KBP English Regular Slot Filling - Comprehensive Training and Evaluation Data 2009-2014	*all pre-2015 data*
LDC2015E47	TAC KBP English Sentiment Slot Filling - Comprehensive Training and Evaluation Data 2013-2014	*all pre-2015 data*
LDC2015E48	TAC KBP English Cold Start - Collected Evaluation Data Sets 2012-2014	*all pre-2015 data*
LDC2015E49	TAC KBP English Surprise Slot Filling - Comprehensive Training and Evaluation Data 2010	*all pre-2015 data*
LDC2015E50	TAC KBP English Temporal Slot Filling - Collected Training and Evaluation Data Sets 2011 and 2013	*all pre-2015 data*

Table 2: 2015 Tri-lingual Entity Discovery & Linking Data

Catalog ID	Title	Size
LDC2015E42	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Knowledge Base	1 knowledge base
LDC2015E43	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Knowledge Base Entries Creation Algorithm	1 algorithm
LDC2015E44	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Pilot Gold Standard Knowledge Base Links V1.1	686 mentions
LDC2015E61	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Pilot Source Corpus	15 documents
LDC2015E75	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Training Data V2.1	30838 mentions
LDC2015E93	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Evaluation Source Corpus V2.0	500 documents
LDC2015E102	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Evaluation Queries V1.2	32,533 queries
LDC2015E103	TAC KBP 2015 Tri-lingual Entity Discovery and Linking Evaluation Gold Standard Entity Mentions & Knowledge Base Links	32,533 mentions

Table 3: 2015 Cold Start Data

Catalog ID	Title	Size
LDC2015E72	TAC KBP 2015 English Cold Start Entity Discovery Sample Data	162 mentions
LDC2015E76	TAC KBP 2015 English Cold Start Evaluation Queries V2.0	2539 queries
LDC2015E77	TAC KBP 2015 English Cold Start Evaluation Source Corpus V2.0	49124 documents
LDC2015E80	TAC KBP 2015 English Cold Start Evaluation Queries and Manual Run	2218 responses
LDC2015E81	TAC KBP 2015 English Cold Start Entity Discovery Evaluation Gold Standard Entity Mentions V1.2	8110 mentions
LDC2015E100	TAC KBP 2015 English Cold Start Evaluation Assessment Results V3.1	30,678 assessments

Table 4: 2015 Event Argument Linking Data

Catalog ID	Title	Size
LDC2015E41	TAC KBP 2015 English Event Argument Linking Training Data	9927 assessments
LDC2015E79	TAC KBP 2015 English Event Argument Linking Evaluation Source Corpus	500 documents
LDC2015E92	TAC KBP 2015 English Event Argument Linking Evaluation Manual Run	5207 arguments
LDC2015E101	TAC KBP 2015 English Event Argument Linking Evaluation Assessment Results V2.0	>7,869 assessments

Table 5: 2015 Event Nugget Data

Catalog ID	Title	Size
LDC2015E73	TAC KBP 2015 Event Nugget Training Annotation	6538 nuggets
LDC2015E94	TAC KBP 2015 Event Nugget and Event Coreference Linking Evaluation Source Corpus	202 documents
LDC2015R26	TAC KBP 2015 Event Nugget and Event Coreference Linking Evaluation Gold Standard Annotation Corpus	6438 nuggets