Overview of Linguistic Resources for the TAC KBP 2016 Evaluations: Methodologies and Results

Joe Ellis, Jeremy Getman, Neil Kuster, Zhiyi Song, Ann Bies, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
{joellis, jgetman, 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 2016, TAC KBP's eighth year of operation, the evaluations focused on five tracks targeting information extraction and question answering technologies: Entity Discovery & Linking, Cold Start, Event Arguments, Event Nuggets, and Belief and Sentiment. Linguistic Data Consortium (LDC) at the University of Pennsylvania has supported TAC KBP since 2009, 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 2016, focusing on changes that were made to meet new requirements.

1 Introduction

In 2016, TAC KBP, an evaluation tracked coordinated by NIST, continued its primary goal of promoting research in automated systems that discover information about entities as found in a large corpus and incorporating this information into a knowledge base. 2016 was the eighth year that TAC KBP was conducted, as well as the eighth year for which Linguistic Data Consortium (LDC) was the primary data provider for the evaluation series, developing and distributing training and evaluation datasets as well as tools and specifications. LDC created a total of 29 new data sets in support of all five tracks for the KBP 2016 evaluations - Entity Discovery & Linking (ED&L), Cold Start (CS), Event Arguments (EA), Event Nuggets (EN), and Belief and Sentiment (BeSt). These data included training and evaluation releases for participants, preliminary releases to coordinators for data previews, and updates to existing releases to improve quality and add new data.

Largely in response to goals for the DARPA DEFT program, which provides funding for TAC KBP data creation, two primary goals emerged during early planning for KBP 2016. The first of these was to increase coordination across the evaluation tracks with an eye toward eventually developing a single, 'all-in-one' track that could test all components of knowledge base (KB) creation and population. The second goal was to increase the number of multilingual evaluation tracks within KBP, specifically those conducted over a mixed collection of English, Chinese, and Spanish source documents. Meeting these goals provided new and interesting challenges to data creation efforts.

This paper describes the processes by which data were developed in support of TAC KBP 2016 as well as the results of those efforts, focusing primarily on new tasks as well as changes to processes as they existed at the end of 2015 in order to meet the goals described above. Sections 2 through 8 discuss the procedures and methodologies for data selection, query development, annotation, and assessment for all TAC KBP data developed in 2016. Section 9 offers concluding remarks. The appendix lists all final datasets released by LDC in support of TAC KBP 2016.

2 Data Selection

LDC assembled separate document collections for a few of the tracks in KBP 2015, which allowed researchers to explore different areas relevant to their own tracks. but produced less layers of annotation over each document. Therefore, as a first step in moving towards greater coordination across tracks, KBP returned in 2016 to the approach of using a single source document collection for all evaluations. From a data development standpoint, this approach has the benefit of producing a greater number of overlapping and complimentary annotations for the same set of source documents, while also reducing the overall number of collections to assemble. However, this approach also requires documents to include a very large number of different features in order to satisfy the diverse set of requirements for all tracks. While the full evaluation corpus includes approximately 90,000 documents, a 505-document, manually-selected subset was used for most of the gold standard data.

The newswire (NW) portion of the 2016 evaluation corpus was selected from a collection of previously unexposed New York Times and Xinhua articles, the former of which was acquired in 2013 and made up of English documents only, and the latter of which was acquired in 2015 and contained Chinese, English and Spanish documents. The discussion forum (DF) part of the source corpus was selected from online threads, which were reviewed by data scouts and later harvested for annotation and distribution. All documents under consideration for use in the evaluations were required to be from within a relatively short, overlapping epoch and at most roughly 800 tokens in length. For DF threads, the latter requirement was met primarily by truncating threads after harvesting.

2.1 Topic Development and Document Selection

In order to help facilitate the selection of documents with a high degree of overlapping entities and events, data scouts search for documents pertaining to a pre-selected set of topics. Topics must pertain to specific, well-defined events of the types annotated in the TAC KBP event tasks (event types are discussed below in section 3). Additionally, topics must be globally newsworthy enough to be discussed in Chinese, English and Spanish documents. Lastly, topics must have the potential to produce documents with ambiguous entities, including synonymous entities (different entities referenced by matching strings), polysemous entities (entities referenced by a variety strings), and entities referenced only by nominal mentions in some documents and only resolving to names in others. For instance, the Rana Plaza collapse in Bangladesh is a specific event with many individual, unnamed victims.

Initial topic selection is performed by senior annotators, who research the productivity of a potential topic in the source newswire pools, record details about which entities and event types are commonly associated with the topic, and then provide an example document containing a representative instance of the topic. Once an initial set of topics is developed, data scouts search for on-topic documents, and tally occurrences of the desired features described earlier. While scouting documents for the 2016 KBP evaluation corpus, over 1,100 documents were reviewed.

As tallies grow sufficiently large, selection of the 505-document core corpus begins. Document selection has to balance multiple needs, including a roughly even balance of genres and languages and sufficient coverage of the 18 event types in TAC KBP, each of which must appear in at least 10-15 documents for each of the 6 language/genre combinations. Related to this, each of a minimum of 50 cross-document event hoppers (event objects including mentions and arguments, described in detail in section 4) has to occur in at least 3 documents, and a minimum of 10 event hoppers each must be mentioned in at least 10 documents. Ambiguous entity mentions also have to be maximized across the corpus.

Scouting of discussion forum sites was done manually using the live web in order to increase variety and thereby maximize topic overlap with the already-obtained NW documents. Some documents selected for inclusion during initial rounds of scouting failed during data processing, and so supplemental documents had to be selected.

Following manual selection of the core source documents, the websites associated with the 250 selected discussion forum threads are harvested and formatted into XML. Afterward, automated selection of the remainder of the 90K-document corpus is performed. This process selects documents using fuzzy name string matching against a list of annotated, named entity mentions, evenly balancing the representation of languages and genres in the final set of selected documents.

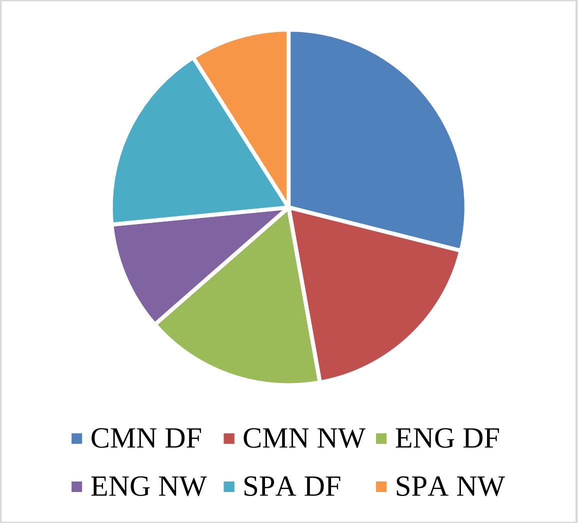

Figure 1: KBP 2016 full corpus token distribution

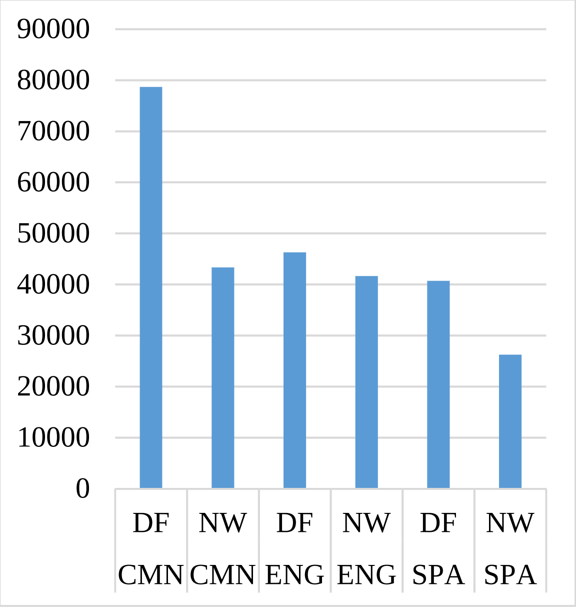

Figure 2: Token counts in the 'core' corpus

3 Entities, Relations, and Events (ERE)

Entities, Relations, and Events (ERE), an annotation task developed by LDC for DARPA's Deep Exploration and Filtering of Text program (DEFT), was first conducted in 2013 with the goal of supporting multiple research directions and technology evaluations. As with earlier related efforts like Automated Content Extraction (ACE) (Doddington et al., 2004; Walker et al., 2006), ERE exhaustively labels entities, relations and events along with their attributes according to specified taxonomies (Song et al., 2015).

As part of the effort to increase coordination across KBP data sets in 2016, ERE annotation was performed as an upstream task in the overall KBP data creation pipeline, providing inputs to downstream annotation tasks supporting ED&L, EA, EN, and BeSt. In looking for ways to free up annotation resources for use elsewhere, using ERE as input was an obvious choice, given that the data largely overlaps with existing entity, relation, and event taxonomies in KBP. Though this approach created some new technical demands and elongated the data pipeline overall (taxing an already-tight timeline), it did away entirely with the need for some previous event annotation tasks, contributed to a significant effort reduction in others (ED&L), and enabled two completely new tasks (cross-document EA and BeSt).

One of the primary motivators for using ERE in TAC KBP is its notion of event "mentions" and event "hoppers", ERE annotation objects which provide convenient definitions for including events in a KB. In ERE, an event mention includes a text extent referencing the event, labels indicating type and other qualities, and usually event arguments – actors and other objects involved in the event. Event hoppers then are clusters of ERE event mentions about the same event that use a more inclusive, less strict notion of event co-reference. Following the guidelines, event mentions can be grouped together into hoppers even when temporal and location arguments are represented at different levels of granularity in the text. For example, an event hopper for an attack event could contain event mentions with the location arguments "Iraq", "Baghdad", and the "Green Zone", despite their differing levels of granularity (Song et al., 2015).

3.1 Changes to ERE Annotation

Some new features had to be added to ERE Entity and Event annotation in order to better meet the needs of the KBP 2016 evaluations and to ease the use of the data by downstream tasks. The most challenging of these extensions to implement was the labeling of individuality as a quality of entities. This label was necessary to support ED&L, which would use entities from ERE as input but considers references to groups of entities as invalid (other ERE entity annotations, such as pronouns, were also filtered out of ED&L inputs). In the definitions and examples below, note the particular difficulty posed by Location (LOC) entities, for which grammatically plural mentions (e.g. “the Hawaiian Islands”) can be considered individuals:

Individual: A unique, single entity in the world, e.g. Barack Obama, the gunman, Rocky Mountains, the biggest lake in North America, Route 66, European Union, Philadelphia, my country

Group: More than one unique, single entity in the world, e.g. the Obama family, multiple victims, America’s mountains, the buildings at Penn, the Soviet Union countries, the Axis powers

Unknown: Can’t tell in context if it is one or more than one unique, single entity in the world (usually applies only to mentions in Chinese), e.g.: 有[人] 在爆炸中受伤。[伤者]被送医（English: People got injured in the bombing. The injured were transported to the hospital).

In addition to the individuality label, the inventory of event types/subtypes was reduced from 9 types and 38 subtypes in 2015 to 8 types and 18 subtypes in 2016 (although ERE data was not used across multiple tracks for KBP 2015, it was used to produce the EN data for that year, and so a shared event inventory was used). The reduced set of event types was jointly selected by DEFT sponsors and program stakeholders and the rest of the organizing committee to better meet stakeholder needs and coordinators’ research interests. Most of the dropped event types and subtypes are scarce in the existing training data (tables 2 and 3 below list the event types used in 2016 and those that were dropped)

Conflict. Attack	Manufacture. Artifact
Conflict. Demonstrate	Movement. TransportArtifact
Contact. Broadcast	Movement. TransportPerson
Contact. Contact	Personnel. Elect
Contact. Correspondence	Personnel. EndPosition
Contact. Meet	Personnel. StartPosition
Justice. ArrestJail	Transaction. Transaction
Life. Die	Transaction. TransferMoney
Life. Injure	Transaction. TransferOwnership

Table 1: ERE event types for 2016

Life. BeBorn	Justice. ChargeIndict
Business. Start	Justice. Sue
Business. End	Justice. Convict
Life. Marry	Justice. Sentence
Life. Divorce	Justice. Fine
Business. MergeOrg	Justice. Execute
Business. DeclareBankruptcy	Justice. Extradite
Personnel. Nominate	Justice. Acquit
Justice. ReleaseParole	Justice. Pardon

Table 2: ERE event types discontinued in 2016

3.2 Results

ERE annotation was performed on all 505 documents that make up the 'core' subset of the 2016 evaluation source corpus. All 18 event types and subtypes, are represented in the data with *Life.Injure* and *Transaction.Transaction* being the least represented and *Contact.Broadcast* and *Conflict.Attack* being the most represented, as shown in figure 1 below.

Early analyses of the data indicate that extracting all event mentions remains a challenge for event annotators. A review of a portion of the data shows variance in annotators interpretation of allowable inference, with generic events being more easily missed, and atomic events faring better than aggregate events. Should ERE be used similarly for data production in TAC KBP 2017, we would at least implement some of the earlier discussed changes to the document selection procedure to allow more time for quality control (QC) of ERE. While improved QC would help catch misses in event annotation, improvement of event detection and extraction overall requires further research.

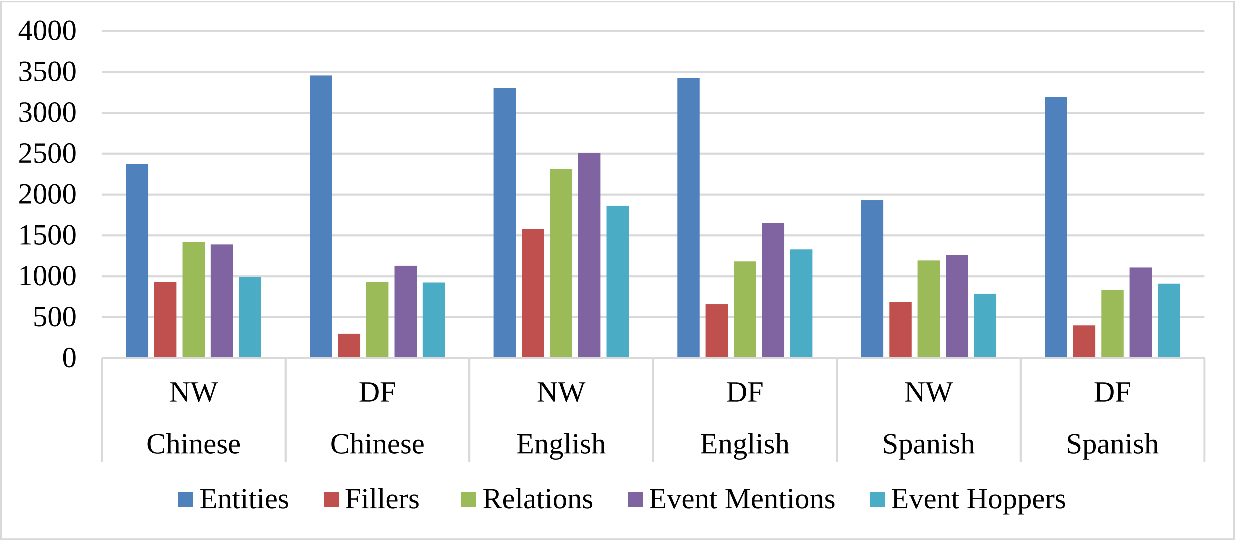

Figure 3: ERE annotation counts in KBP 2016 evaluation data

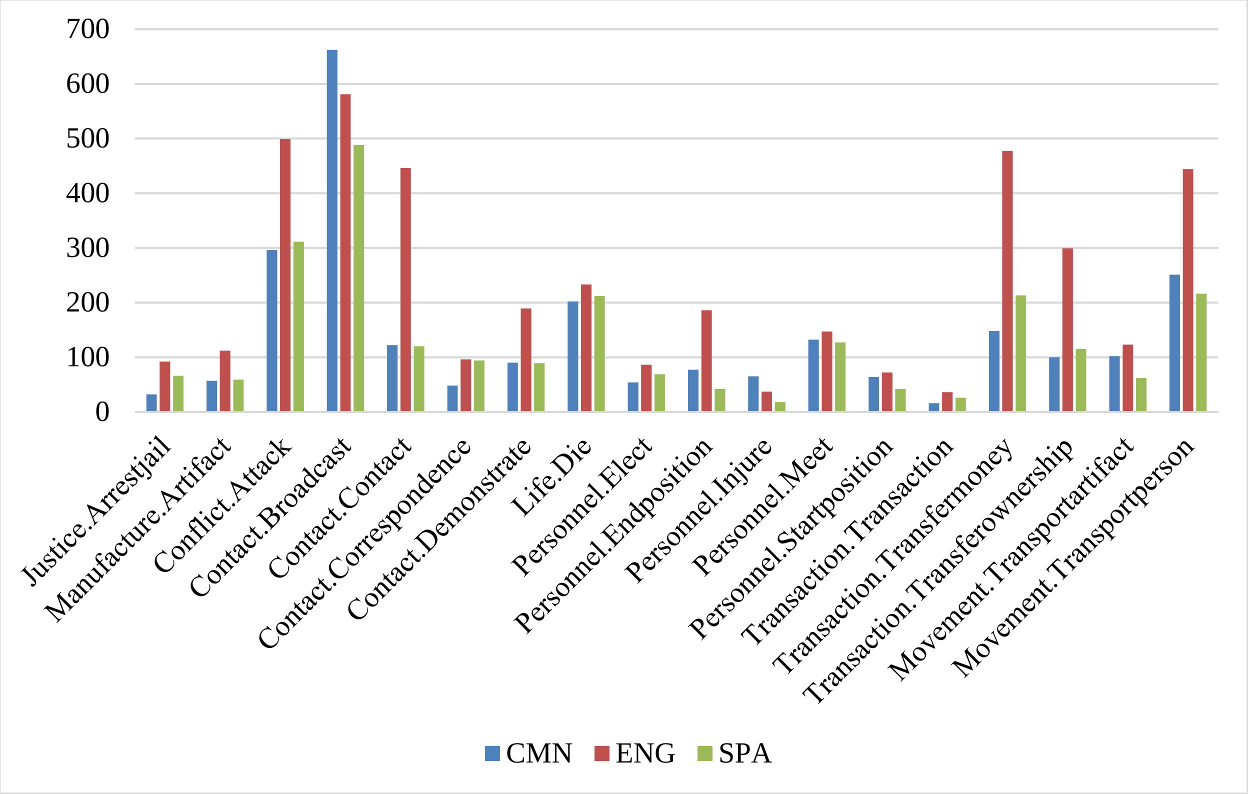

Figure 4: ERE event mentions in KBP 2016 evaluation data

4 Entity Discovery & Linking

The goal and overall approach to data creation in 2016 for Entity Discovery & Linking (ED&L) remained relatively consistent with the approach used in 2015. That is, ED&L annotation in 2016 consisted of exhaustive entity extraction and cross-document clustering from a cross-lingual collection of documents, as well as linking of entities to an external KB.

That said, behind-the-scenes changes to ED&L in order to meet the goal of increased coordination of data did impact data creation efforts. This was primarily seen in the importation of entities from ERE, which shifted the focus of the entity discovery part of the task from exhaustive annotation to quality control over the imported data, reducing time spent on the task overall. Also as part of the effort of increasing coordination, nominal entity mentions were added for all entity types in 2016. One last important change to the task was the mode by which annotators searched the KB to which entities were linked, which was motivated by the knowledge that the resource used for this purpose in 2015 was unsustainable.

4.1 Changes to Gold Standard Data Development in 2016

Rather than starting with a blank slate, as had been done since the task was first conducted in 2014, ED&L annotators started by reviewing all of the entity mentions and equivalence class clusters that were imported from ERE. All imported mentions were highlighted in the source documents displayed to annotators so that they could check for extent errors, mentions imported that might be at variance with the ED&L guidelines (though possibly correct fore ERE), and outright misses. Additionally, equivalence class clusters were grouped together next to the document so that annotators could review and adjust if necessary.

As kits were completed, a comparison script identified changes made by ED&L annotators by reporting on mismatches between ERE and ED&L annotations. Such mismatches were thoroughly reviewed, to ensure that variance only occurred in cases for which there were clear errors in the ERE data. During these reviews, three general categories of changes emerged, namely, ED&L entity mentions that (a) had extent offsets which were incongruent with but overlapped with offsets of an ERE mention, (b) matched an ERE text extent but was at variance with one or more labels (mention type, entity type, or specificity), and (c) were true misses – entity mentions completely absent from the ERE data.

As mentioned earlier, changes were necessary in 2016 to the mechanism by which ED&L annotators searched the KB for nodes to which entities should be linked. In 2015, BaseKB was selected for use as the official KB for linking. While LDC was able to produce a human-readable version of the RDF triples in BaseKB for annotators to review, search results were poor and workarounds had to be employed.

Researchers at NIST took on the challenge of developing a new search engine for getting results out of BaseKB and LDC assisted in the process by testing and providing examples of problematic results. While significant performance gains in searching were made compared to the results seen in 2015, there were still some known issues in results rankings at the point when ED&L data production had to begin in order to finish before the evaluation. As such, a few workarounds were produced for finding BaseKB mIDs, including a lookup table of known problematic entities and Wikidata (www.wikidata.org/wiki/Wikidata:Main_Page), which acquired much of the content of Freebase (a superset of BaseKB) before the resource went offline.

One other important changes to ED&L this year to increase coordination across data was the expansion of the validity of nominal mentions. In 2015, nominal had been relegated to only English-language person entities. However, for 2016, heads of individual nominal mentions were annotated exhaustively across all three languages and all five entity types. Two other changes made to ED&L annotation this year were the removal of Title and embedded (intra-token) mentions.

4.2 Results

While importing ERE entities helped to decrease ED&L annotation rates overall, data developers did spend more time correcting entity mention inputs than was desired or anticipated. Thus, further work is needed to allow annotator cycles for additional review of ERE both for quality control and prior to its incorporation into the ED&L pipeline. That said, as mentioned earlier, the timeline for ERE annotation was compressed due to complications in harvesting documents during data selection. Therefore, in addition to further addressing ERE-ED&L guidelines mismatches, we believe that by simply wrapping up data selection earlier and, thereby, giving ERE more time for quality control, would further improve ED&L annotation rates.

Additionally, the new KB search process, while clearly an improvement over the 2015 setup, still required annotators to use workarounds for finding mIDs in the KB, which is error prone and slows annotation. As a result of the testing of KB search results conducted before data production began, we know of at least 20 entities that are in the KB but difficult to find because they consistently received low rankings in search results. As such, there is certainly some number of other, unknown entities for which search results rankings were poor, meaning that there could be a higher number of linking errors in the data compared to previous years. Therefore, additional improvements to search results rankings would improve efficiency and errors considerably and this remains an area for further work.

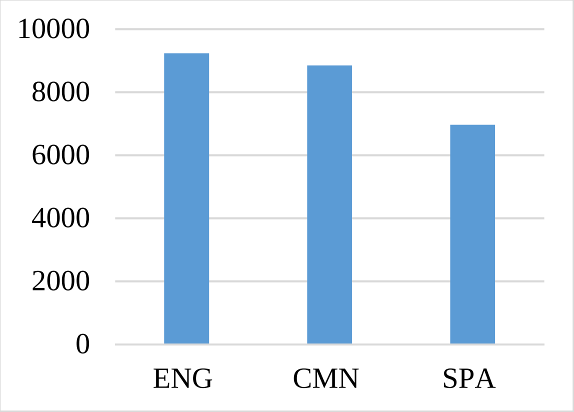

Figure 5: ED&L 2016 entity mentions

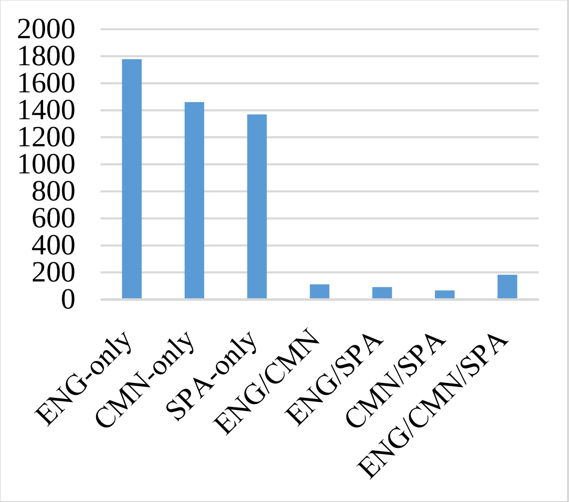

Figure 6: ED&L 2016 entity clusters

5 Event Argument

From a data development perspective, the needs of the 2016 Event Argument (EA) task were completely different from those supporting EA in 2014 and 2015. In 2014-2015, LDC created an event argument manual run over the documents in the EA source corpus, which consisted of annotating all of the unique event arguments that occurred in that corpus, and, in 2015, grouping those event arguments into event hoppers. Following the evaluation window, LDC annotators performed argument-level assessment of all event arguments produced by participant systems and LDC annotators, and, in 2015, as with the manual run, annotators then grouped the (correct and inexact) arguments into event hoppers.

In 2016, however, instead of an assessment paradigm, LDC, with the help of BBN, created a set of gold standard EA annotations. The gold standard was a modified/expanded version of the ERE data developed from the core docs in the evaluation source corpus, and the system submissions were scored against this gold standard, instead of through argument-level assessment. In addition to the switch to a gold standard paradigm, an English-only cross-document task was added, for which LDC selected queries, produced a set of manual responses, and performed assessment.

5.1 Gold Standard Data Development

Following ERE data development, the annotations were augmented by running a script developed by BBN over the data and then having ERE annotators review the results for validity. The purpose of the augmentation pass was to add inferred arguments that are invalid following ERE guidelines and difficult for human annotators to find in general. In large part this translated to arguments that could be inferred by locational containment. For example, a *Conflict.Attack* event that had Baghdad annotated as the Place of the event might have Iraq added as an additional, inferred Place during the augmentation pass.

5.2 Cross-Document Query Selection & Manual Run

To support the new cross-document component of EA, annotators selected queries comprised of a single event argument pertaining to an event hopper in the gold standard EA annotations described above. Given the anticipated difficulty of the task for systems, potential queries included only events for which a named event argument had been annotated, were sourced only from English documents (thus the task was English-only in its first year), and excluded the 3 new event types added to EA in 2016 (*Contact.Contact*, *Contact.Broadcast*, and *Transaction.Transaction*). Annotators also were instructed to limit potential queries to event arguments that indicated relatively simple, low-granularity event hoppers.

In addition to providing "easy" queries as described above, the final set of queries had to represent roughly equally the 15 event types in scope for the cross-document task. Queries were also required to be productive, with the event indicated by the query occurring in at least 5-10 documents in the English portion of the source corpus. Some less-productive queries were included as well, however, in order to ensure that rarer or more difficult event types were represented in the query set. In total, 51 unique English queries were selected.

LDC also produced an exhaustive manual run for queries, which was performed over the entirety of the 30K-document English portion of the TAC KBP evaluation source corpus. A response for the manual run consisted of justification strings containing whatever portion or portions of a document was needed to prove that the event indicated by the relevant query occurred in the given document. A document could be returned more than once, if each instance was in response to a different query.

After the cross document EA evaluation was conducted, it was discovered that, despite the efforts taken to produce relatively simple queries, systems were largely unsuccessful in finding the entry points indicated by the queries. As such, an additional 249 “derived” queries were produced from systems’ responses by BBN in an effort to better measure precision given low system recall. For these 249 queries, LDC did not produce a cross-doc manual run.

5.3 Cross-Document Assessment

For the assessment portion of cross-document EA, annotators reviewed all of the responses to both the 51 queries manually selected by LDC and the 249 derived queries generated from system responses. While the assessment pools for the manual queries included responses from both LDC and systems, the pools for the derived queries included only responses from systems. In total, assessors reviewed 1,221 responses to the 51 manual queries and 6,476 responses to the automated queries.

During assessment, assessors reviewed each response individually, and decided whether or not the response’s justification string(s) proved that a document contained an instance of the event indicated by the relevant query. If the assessor determined that the response did indeed reference the same event, it was marked CORRECT. If the response was determined to contain an event of the same type as the query event, but not the query event itself, the response was marked ET_MATCH (event type match). If the response was judged to contain neither the query event nor some other event of the same type, the response was marked WRONG.

5.4 Results

As mentioned earlier, the human-produced, cross-document EAL queries produced a relatively low number of system responses, prompting the need to generate a set of derived queries in order to better measure precision. However, it should be noted that query developers struggled in selecting queries within the parameters of simplicity set forth by coordinators, reviewing over 1,300 potential queries before setting on the final set of 51 used in the evaluation. These two facts will be difficult to balance should a similar approach to measuring cross-document event detection be used in the future.

Although scores have not been made available at the time of writing, table 4 below gives a breakdown of how responses were assessed. While nearly all (98%) of LDC’s responses were marked correct, only 17% of system responses were judged as such. Encouragingly, however, most system responses were not entirely wrong; 74% pointed to events matching the query events in type.

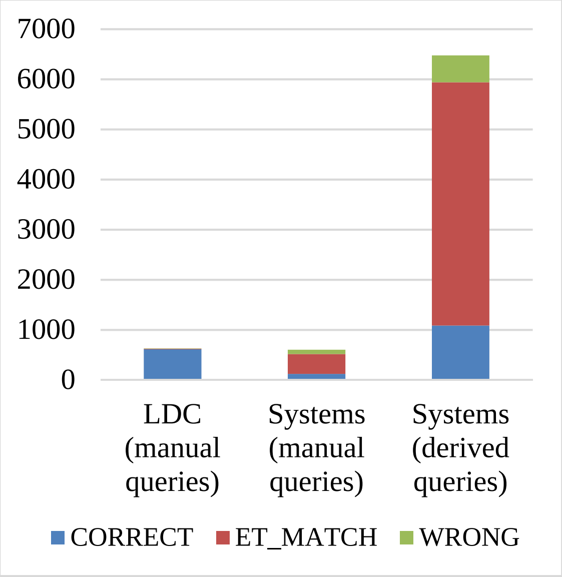

Figure 7: 2016 Cross-document EA assessments

6 Event Nugget and Linking

Contrary to EA, the Event Nugget and Linking track (ENL) changed very little in 2016 as compared to the previous year. As in 2015, there was no separate annotation task conducted solely to support the ENL evaluation in 2016; the data are entirely produced by running a script over ERE data to extract and reformat a subset for use by ENL. The only change made to the task in 2016 was the use of Chinese, Spanish, and English source documents as inputs as the data had been English-only in previous iterations.

An event ‘nugget’, as defined by the task, includes a text extent, a classification of event type and subtype, and an indication of whether realis mood was used to describe the event (Ellis et al., 2015), all of which are defined following ERE specifications. Similarly, event nugget linking uses the ERE event hopper concept (described in section 3) to define the approach to clustering event nuggets.

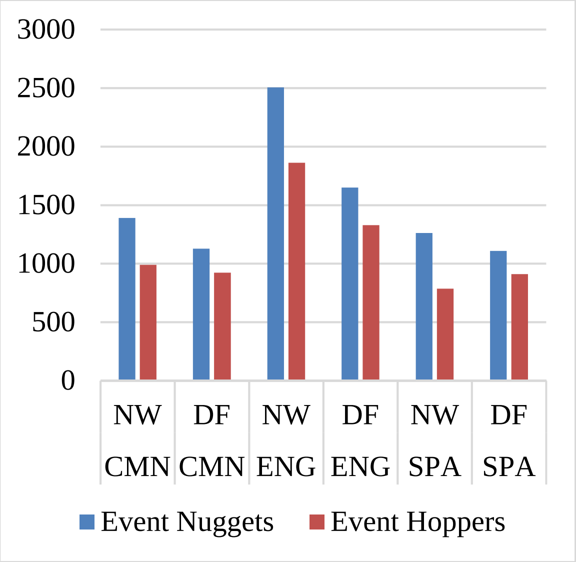

Figure 8: 2016 EN nuggets and hoppers

7 Belief and Sentiment

Belief and Sentiment (BeSt) is a new track for TAC KBP in 2016. It emerged as a task from DARPA’s DEFT program, with the goal of augmenting information about entities, relations, and events in a knowledge base with beliefs and sentiment. Until this year, the annotation of belief for the DEFT program included exhaustive annotation of all propositions in a document with respect to the speaker/writer’s level of committed belief in the proposition, and a pilot evaluation of belief detection systems was held within DEFT in December 2014. In order to facilitate both the addition of sentiment to the task and the connection of both beliefs and sentiment to objects in a KB, the new BeSt track was developed for TAC KBP 2016.

BeSt requires that belief and sentiment be annotated with respect to entities, relations, and events as annotated in ERE. Entities can be holders (or reporters) of belief and/or sentiment; relations and events can be the targets of belief and/or sentiment. Entities can also be targets of sentiment. The BeSt annotation also labels an entity's role in an event as a target of belief, separate from belief in the event itself, but this part of the annotation was not evaluated in 2016.

7.1 Annotation Procedure

Input to the BeSt annotation task is an ERE-annotated document. A single annotator performs two passes over the list of ERE annotations: one for belief, and one for sentiment. For belief, all possible targets are marked with one of the following belief type labels. The term "proposition" in the descriptions below refers to the existence of the relation or event and/or the role of entities as event arguments (note that unrealistically simple examples below are used only to clarify differences in the definitions).

Committed Belief (CB) -- the holder believes the proposition with certainty
Example: "John traveled to Turkey"
The writer asserts with certainty that the "travel" event occurred.

Non-committed Belief (NCB) -- the holder believes the proposition to be possibly, but not necessarily, true
Example: "I think John traveled to Turkey"
The writer indicates some uncertainty that the "travel" event occurred.

Reported Belief (ROB) -- the holder reports the belief as belonging to someone else, without specifying their own belief or lack of belief in the proposition
Example: "Mary says John traveled to Turkey"
The writer attributes certainty that the "travel" event occurred to Mary but does not indicate her or his own belief.

Not Applicable (NA) -- the holder expresses some cognitive attitude other than belief toward the proposition, such as desire, intention, or obligation.
Example: "I hope John travels to Turkey"
The writer indicates no belief about whether the "travel" event will occur, only a wish that it should.

For relations, the annotator treats the entire relation as a whole and does not separate belief in an entity's participation in the relation from belief in the relation itself. However, for events as targets of belief, the annotator does provide a separate judgment about whether the holder believes in each entity-argument's role in the event as well as the event itself. For example, given the text extent: "*ISIS may have been responsible for the bombing*", annotators would indicate that the writer expresses a committed belief that the "bombing" event occurred, but a non-committed belief about the role of ISIS as the agent of the bombing. Beliefs about entities' roles in events were not evaluated in 2016, but they do appear in both the training and gold standard evaluation data.

In addition to the target and belief-type, the holder of the belief is explicitly indicated (and in the case of reported belief, a chain of attribution is annotated), and the polarity of the belief is indicated. Positive polarity means belief that the event/relation/entity-participation did occur, while negative polarity means belief that it did not occur. So "*Paris is in France*" would be labeled as a committed belief with positive polarity, while “*Paris is not in France*” would be labeled as a committed belief with negative polarity.

Sentiment is annotated with entities (independent of their role in an event or relation), relations, and events as targets. Polarity indicates positive or negative sentiment, and holder (including chain of attribution where relevant) is indicated as in belief annotation.

The sarcasm attribute signals whether the polarity of the belief and sentiment was tagged as the opposite of what a literal reading of the text (without context) would suggest. The sarcasm flag was not evaluated in 2016 but does appear in the training and evaluation gold standard annotation.

The targets and holders of belief and/or sentiment are entity, relation, and event mentions annotated in ERE. Beliefs and sentiments toward other targets are not annotated.

Once the first-pass annotator has completed annotation of both sentiment and belief on a document, all documents in the evaluation set were put through a second pass, in which a senior annotator reviewed the annotations, with a particular focus on sentiment, since lower consistency for sentiment was identified during annotation of the training data.

7.2 Results

For the 2016 BeSt evaluation, LDC produced training data as well as gold standard annotation for the evaluation set. The evaluation data was the 505-document core set of ERE annotated data described earlier. The tables below show quantities of training and evaluation data as well as a breakdown of annotations produced

Differences in total number of belief and sentiment annotations across languages in the training data is a result of the fact that there was a larger amount of data annotated in English than in the other two languages, with Spanish having the smallest data volume. But in the evaluation set, each language had approximately the same volume of data, yet English had a higher total number of annotations. This seems to have been due to the English newswire data being particularly rich with ERE annotations and therefore having more targets for belief and sentiment.

For all three languages, there are some differences in distribution of belief types between the training and evaluation data (relatively fewer “NA” beliefs across all three languages, for example). A likely explanation for this mismatch is the fact that the evaluation data was approximately half newswire, while the training data was primarily discussion forum.

Perhaps surprisingly, the increased proportion of newswire data in the evaluation set does not seem to have affected the presence of sentiment, except perhaps in English..

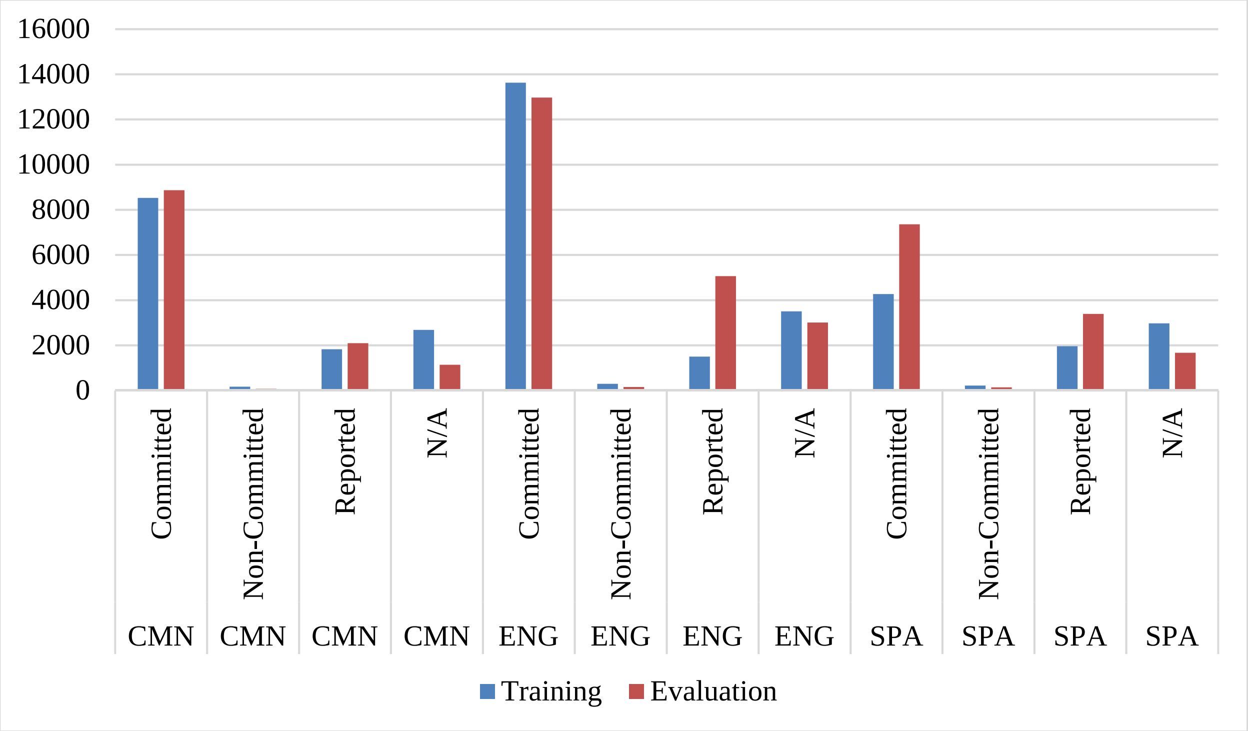

Figure 9: 2016 BeSt belief type annotations

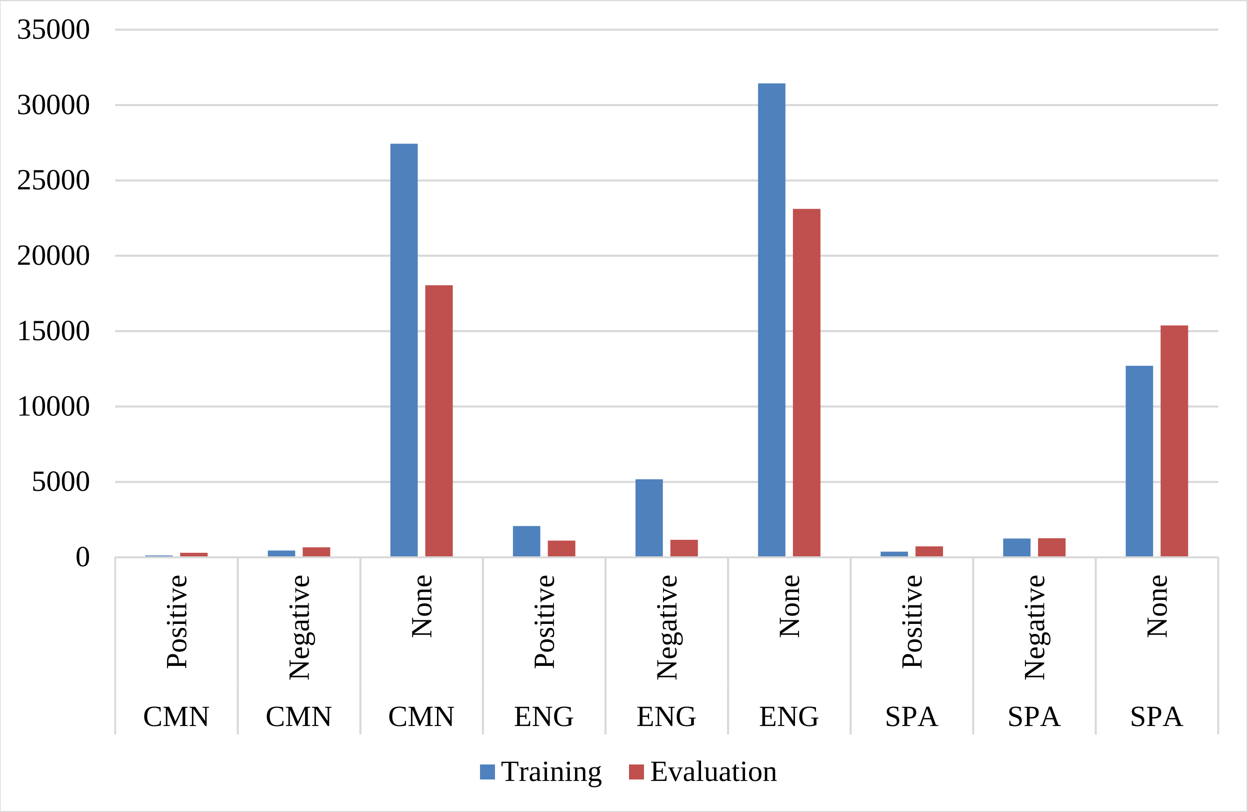

Figure 10: 2016 BeSt sentiment type annotations

8 Cold Start

At a very high level, data development in support of Cold Start for 2016 was relatively consistent with the approach used in 2015. That is, annotators created a set of queries intended to navigate and evaluate system-submitted KBs, a "manual run" of human-produced responses to the queries, and assessments for a subset of responses produced during the evaluation.

That said, however, lower-level implementation and procedural changes were necessary in 2016 in order to support both the expansion of Cold Start from a mono- to cross-lingual task as well as the addition of nominal entity mentions as valid responses to queries. Other changes to Cold Start data development were taken to address low recall seen in the 2015 manual run and other lessons learned in the previous year.

Unlike all the other tracks discussed up to this point, Cold Start is the only TAC KBP 2016 data that did not directly utilize ERE data as input. This decision was primarily made due to timeline restrictions – since the insertion of ERE data development into the pipeline was lengthening an already-tight timeline, we wanted to allow for at least one data set to be developed concurrently with ERE. Cold Start was an obvious choice for this purpose because, since queries and responses for the manual run were to come from across the full 90K corpus, there were less advantages to be had from the ERE data since it was restricted to the manually-selected subset.

8.1 Changes to Query and Manual Run Development

The results of 2015 data development indicated two problems with the approach to query development that had been taken that year. Primarily, recall for the manual run was low – 19% – and our hypothesis was that merging query development and manual run development (a move that had been taken to reduce effort overall) was causing annotators to focus more on finding interesting queries rather than finding all responses to each of those queries. As such, the two tasks were largely separated this year (some initial response generation is necessary during query development in order to count productive queries – those with valid responses in the corpus – and to determine whether those queries could produce responses from English, Chinese, and/or Spanish sources).

Another problem in the previous Cold Start query development approach centered around null queries – those that are not known to contain valid responses in the corpus. In 2015, LDC and other KBP coordinators decided to automate the development of null queries by copying productive queries and replacing the slots that were used in them. The theory was that, while this approach would not guarantee that the queries were truly null, response counts for them would still be relatively low. However, this theory proved to be false; many responses were produced for null queries in 2015. This complicated the selection of responses for assessment and so, for 2016, null queries were developed manually, ensuring that most if not all were truly null.

The final, and perhaps most consequential, changes to query development in 2016 were taken in order to support the production of multi-lingual queries. In 2015 and 2016, annotators generate Cold Start queries via kits centered around 1-5 mentions of a single query (or ‘entry point’) entity, which is then paired with sets of 1-2 slots (1 slot for a 1-hop query, 2 slots for a 2-hop query) to arrive at a number of queries each starting with the same entry point entity. When Cold Start was mono-lingual, this process was relatively short as each kit needed only to be reviewed by 2 people – a single annotator generated a kit with a set of queries (first pass), which was subsequently reviewed by another, usually more senior, annotator (second pass). Kits for multilingual queries, however, require review by up to 6 annotators – one for each language in the first pass, and one for each language in the second pass. Note though that sometimes less than 3 reviews were conducted in the second pass, especially as annotators became more experienced with the task.

As indicated earlier, after the set of queries was finalized, annotators began production on the manual run. Picking up where query development ended, annotators were given query development kits in essentially the same state as they had been in at the end of that task. However, entry point entities and queries could no longer be edited (or added or deleted) and annotators were simply tasked with going through each of the existing queries and adding all responses they could find, spending no more than 1 hour per kit on first passes. Note that, like the query development task, each kit in the manual run required review by up to 6 annotators – one for each language in the first pass, and up to one for each language in the second pass.

Total queries	1,077
Total productive queries	915
Total entry-point entities	208
Total manual run responses	4,739

Table 3: 2016 Cold Start data volumes

8.2 Changes to Assessment

Like query development and manual run production, the overall approach to Cold Start assessment was relatively consistent with that taken in previous years. Assessors were presented with a set of responses for a given query and had to determine the validity of fillers and justification for each. Afterward, responses marked as correct or inexact were co-referenced in order to indicate redundant responses as well as the total number of correct responses for each query.

One of the changes to assessment is of course the same as what was described above for query development and manual run production. In order to deal with the fact that responses for a given query might be in English, Chinese, or Spanish, each kit had to be reviewed by up to 6 different assessors - one for each language in the first pass, and up to one for each language in the second pass.

However, the addition of nominal entity mentions as valid responses proved to have an even greater effect on the assessment process. First, assessors added a label to each correct and inexact entity-type response in order to indicate whether the reference was a named or nominal mention. Additionally, following the final round of quality control, for any clusters of correct and inexact responses for which only nominal mentions of the filler entity were returned, annotators searched the corpus to determine whether a named mention of the entity existed anywhere in the source collection, adding such strings to the data if found.

8.3 Results

Results were mixed for LDC's manual run in 2016 as compared to the previous year. For English, the only language for which a comparison can be made, we were successful in increasing recall, with a 15% gain over 2015 results. However, there was a slight drop (1%) in precision and so further analysis will be necessary to determine the causes of human errors, especially to ascertain whether changes to the data development approach may have been the cause.

Year	Lang	Precision	Recall	F1
2015	ENG	81%	19%	30%
2016	ENG	80%	34%	48%
2016	CMN	76%	25%	38%
2016	SPA	87%	64%	74%
2016	Cross-lingual	78%	35%	48%

Table 4: LDC's scores for Cold Start

9 Conclusion

This paper discussed the linguistic resources produced in support of the TAC KBP 2016 evaluations, focusing on modifications to the data creation processes, descriptions of the datasets, and analysis of how results compared to previous efforts. Future work will include incorporating lessons learned into processes that will be used again in the future, developing entirely new procedures to accommodate new research goals in the future, and repackaging and updating documentation for data created this year so that it will be more readily useable in the future by system developers, especially 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.

10 References

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel. (2004). Automatic Content Extraction (ACE) Program - Task Definitions and Performance Measures. *4th International Conference on Language Resources and Evaluation* (LREC 2004), Lisbon, May 24-30.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, Stephanie M. Strassel. 2015. Overview of Linguistic Resources for the TAC KBP 2015 Evaluations: Methodologies and Results. *TAC KBP Workshop 2015*: National Institute of Standards and Technology, Gaithersburg, MD, November 16-17.

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).

Christopher Walker, Stephanie Strassel, Julie Medero and Kazuaki Maeda. 2006. *ACE 2005 Multilingual Training Corpus*. Linguistic Data Consortium, LDC Catalog No.: LDC2006T06.

Appendix A: Data Available to KBP Performers in 2015

Table 1: 2009 – 2015 TAC KBP Data Sets Condensed

Catalog ID	**Title**	**Release Date**
LDC2014T16	TAC KBP Reference Knowledge Base	*all pre-2016 data*
LDC2015E17	TAC KBP Chinese Entity Linking Comprehensive Training and Evaluation Data 2011 - 2014	*all pre-2016 data*
LDC2015E18	TAC KBP Spanish Entity Linking - Comprehensive Training and Evaluation Data 2012 - 2014	*all pre-2016 data*
LDC2015E19	TAC KBP English Entity Linking - Comprehensive Training and Evaluation Data 2009 - 2013	*all pre-2016 data*
LDC2015E20	TAC KBP English Entity Discovery and Linking - Comprehensive Training and Evaluation Data 2014	*all pre-2016 data*
LDC2016E38	TAC KBP English Event Argument Extraction - Comprehensive Pilot and Evaluation Data 2014 - 2015	*all pre-2016 data*
LDC2015E45	TAC KBP Comprehensive English Source Corpora 2009-2014	*all pre-2016 data*
LDC2015E46	TAC KBP English Regular Slot Filling - Comprehensive Training and Evaluation Data 2009-2014	*all pre-2016 data*
LDC2015E47	TAC KBP English Sentiment Slot Filling - Comprehensive Training and Evaluation Data 2013-2014	*all pre-2016 data*
LDC2016E39	TAC KBP English Cold Start - Collected Evaluation Data Sets 2012-2015	*all pre-2016 data*
LDC2015E49	TAC KBP English Surprise Slot Filling - Comprehensive Training and Evaluation Data 2010	*all pre-2016 data*
LDC2015E50	TAC KBP English Temporal Slot Filling - Collected Training and Evaluation Data Sets 2011 and 2013	*all pre-2016 data*
LDC2016E36	TAC KBP English Event Nugget Detection and Coreference - Comprehensive Training and Evaluation Data 2014-2015	*all pre-2016 data*
LDC2016E36	TAC KBP English Event Nugget Detection and Coreference - Comprehensive Training and Evaluation Data 2014-2015	*all pre-2016 data*
LDC2016E36	TAC KBP English Event Nugget Detection and Coreference - Comprehensive Training and Evaluation Data 2014-2015	*all pre-2016 data*

Table 2: 2016 TAC KBP Data

Track	**Catalog ID**	**Title**
All	LDC2016E63	TAC KBP 2016 Evaluation Source Corpus V1.1
All	LDC2016E64	TAC KBP 2016 Evaluation Core Source Corpus
BEST	LDC2016E71	TAC KBP 2016 Eval Core Set Rich ERE Annotation
CS	LDC2016E65	TAC KBP 2016 Cold Start Evaluation Queries
CS	LDC2016E69	TAC KBP 2016 Cold Start Evaluation Queries and Manual Run
CS	LDC2016E106	TAC KBP 2016 Cold Start Evaluation Assessment Results V2.0
EA	LDC2016E49	TAC KBP 2016 English Event Argument Linking Pilot Source Corpus
EA	LDC2016E51	TAC KBP 2016 English Event Argument Linking Pilot Queries and Manual Run
EA	LDC2016E60	TAC KBP 2016 English Event Argument Linking Pilot Gold Standard
EA	LDC2016E59	TAC KBP 2016 English Event Argument Linking Pilot Assessment Results
EA	LDC2016E74	TAC KBP 2016 English Event Argument Linking Evaluation Queries and Manual Run
EA	LDC2016E73	TAC KBP 2016 Eval Core Set Rich ERE Annotation with Augmented Event Argument V2
EA	LDC2016E107	TAC KBP 2016 English Event Argument Linking Evaluation Assessment Results V2.0
ED&L	LDC2016E68	TAC KBP 2016 Entity Discovery & Linking Evaluation Gold Standard Entity Mentions and Knowledge Base Links
EN	LDC2016E67	TAC KBP English Event Nugget Training Data - Character Based Format Conversion
EN	LDC2016E72	TAC KBP 2016 Eval Core Set Event Nugget Annotation