Overview of Linguistic Resources for the TAC KBP 2017 Evaluations: Methodologies and Results

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
{jgetman, joellis, zhiyi, garjen, 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 2017, TAC KBP's ninth 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 2017.

1 Introduction

In 2017, TAC KBP, a set of evaluation tracks coordinated by NIST, continued its primary goal of promoting research in automated systems that discover information about entities as found in a large corpus of unstructured text and populating this information into a knowledge base. Linguistic Data Consortium (LDC) was the primary data provider for the evaluation series in 2017, the ninth year in which TAC KBP was conducted. To this end, LDC created a total of 21 new data sets in support of the five tracks making up the KBP 2017 evaluations - Entity Discovery & Linking (ED&L), Cold Start (CS), Event Arguments (EA), Event Nuggets (EN), and Belief and Sentiment (BeSt).

By design, these five evaluation tracks were the same as those making up the 2016 KBP evaluations. As such, resource creation requirements remained largely the same as those in 2016, as did the methods utilized to meet those requirements. There were, however, some key differences between the 2016 and 2017 versions of certain tracks, most notably Cold Start, which was expanded to include extraction of events and sentiment in addition to relations, which were previously the sole focus of the Cold Start track. The data produced by LDC in 2017 included new test sets for all evaluation tracks, as well as improved versions of previous years' data sets for participants to use as training and development data .

This paper describes the processes by which data were developed in support of TAC KBP 2017 as well as the results of those efforts, focusing primarily on changes to processes as they existed at the end of 2016 in order to meet the goals described above. Sections 2 through 5 discuss the procedures and methodologies for data selection, query development, annotation, and assessment for

all TAC KBP data developed in 2017. Section 6 offers concluding remarks. The appendix lists the datasets released by LDC in support of TAC KBP 2017.

2 Data Selection

For 2017, KBP continued the approach taken in 2016 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. The full evaluation corpus includes approximately 90,000 documents, selected from LDC's existing newswire and discussion forum[1] data archives. A single, manually selected subset of 500 documents was used for all tasks with gold standard data (referred to as the "core" set), while assessment of system responses for Cold Start could include documents from the full 90,000-document evaluation set.

The newswire (NW) portion of the 2017 evaluation corpus was selected from a collection of previously unexposed New York Times English data originally collected by LDC in 2013, and Xinhua Chinese, English and Spanish data collected in 2015. The discussion forum (DF) part of the source corpus was selected from threads originally collected by LDC in 2016. All documents under consideration for use in the evaluations were required to be from within a specified epoch and at most roughly 800 words in length. For DF threads, the latter requirement was met primarily by truncating threads after harvesting.

2.1 Topic Development and Document Selection

As was done in 2016, in order to help facilitate the selection of documents with a high degree of overlapping entities and events, annotators reviewed the newswire and discussion forum source data collection to select 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. 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.

Initial topic selection is performed by senior annotators, who research the productivity of a potential topic in the newswire collection, record details about which entities and event types are commonly associated with the topic, and then select an example document containing a representative instance of the topic. Once an initial set of topics is

[1] Discussion forums contain threaded discussions with multiple posts by different authors, and are informal and interactive in style, often including considerable amounts of non-standard grammar and spelling.

developed, annotators search the whole corpus for relevant documents and tally occurrences of the desired features described earlier. While scouting documents for the 2017 KBP evaluation corpus, over 1,000 documents were reviewed.

As tallies grow sufficiently large, selection of the 500-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. Ambiguous entity mentions also have to be maximized across the corpus. Table 1 shows the distribution of genres and languages for the core corpus.

Lang	genre	doc	words
Cmn	NW	83	33,683
Cmn	DF	84	49,932
Eng	NW	83	32,572
Eng	DF	84	42,891
Spa	NW	83	29,615
Spa	DF	83	42,850
Total		500	231,543

Table 1: 2017 Core Corpus

Following manual selection of the core source documents, 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 manually labeled named entity mentions, evenly balancing the representation of languages and genres in the final set of selected documents.

3 Entities, Relations, and Events (ERE)

Since 2016, Rich ERE annotation has been integrated into the TAC KBP evaluations as an upstream task in the overall KBP data creation pipeline. ERE includes the annotation of entities, relations, and events and their attributes, according to a specific taxonomy, and these annotations then become input to the downstream gold standard KBP annotations. The entities from ERE feed into ED&L annotation, and event arguments, nuggets and hoppers are extracted from ERE to support EA and EN evaluations. In addition, BeSt annotation uses the full ERE annotation as input to provide the targets of belief and sentiment. In order to better meet the needs of the KBP evaluation, two changes were made to ERE, which was developed by LDC for DARPA's Deep Exploration and Filtering of Text (DEFT) program. First, for entities, we added labeling of individuality. Second, for events, the inventory of event types/subtypes was reduced from 9 types and 38 subtypes in the training data to 8 types and 18 subtypes in the evaluation data.

Table 2 shows the total volume of ERE annotation produced in support of the TAC KBP 2017 evaluations.

	Genre	**English**	**Chinese**	**Spanish**
2017 Eval	NW	33Kw	34Kw	30Kw
	DF	43Kw	50Kw	43Kw

Table 2: ERE Data Volumes

3.1 Entity Discovery & Linking

The goal and overall approach to data creation in 2017 for Entity Discovery & Linking (ED&L) remained relatively consistent with the approach used in 2016. That is, ED&L annotation in 2017 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.

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 that might be at variance with the ED&L guidelines (though possibly correct fore ERE), and outright misses.

As documents were completed, an automatic process reported changes made by ED&L annotators resulting in mismatches between ERE and ED&L annotations. Such mismatches were thoroughly reviewed, to ensure that changes were made only 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.

After ED&L annotators had finished reviewing all entity mentions imported from ERE, each finalized equivalence class cluster was then linked to a node in the KB or marked as NIL (indicating that the entity did not have a node in the KB). Table 3 shows the number of entity mentions that were either linked to the KB or marked as NIL for each language.

Status	**CMN**	**ENG**	**SPA**
Linked to KB	7,673	4,572	4,982
NIL	2,573	2,343	2,230
Total	10,246	6,915	7,212

Table 3: ED&L 2017 entity mentions

3.2 Event Nugget and Coreference

In 2017, as in 2016, the EN data was produced automatically, by running a script over the ERE annotations on the core set of 500 documents. The resulting output was extracted and reformatted for use by EN (Ellis et al, 2016).

3.3 Event Argument

In 2017, as in 2016, LDC created a set of gold standard EA annotations based on event annotation in rich ERE. In 2016, event arguments were augmented based on a script provided by BBN that was followed by manual validation of the automatically augmented event arguments. In order to facilitate a more exhaustive augmentation pass in 2017, instead of relying on automatic augmentation, LDC performed manual event argument augmentation to add arguments that were considered valid for the event argument annotation scheme, but not for Rich ERE, according to the following guidelines:

- If X fills a Place, Origin, or Destination argument role and there exists a Y in the document such that Y contains X (either based on context or world knowledge), then a new argument is created for the same event mention with the same role that is filled by Y.
- If X is an entity mention of type GPE that either fills an Agent role or

modifies the entity mention that fills an Agent role in a Personnel event, then X should also be added as a Place argument for that event.

- For movement.transportperson, if X is an entity mention that could potentially fill both the Agent AND the Person argument roles, ERE tags it only as Agent, and X should be added as the Person argument in event argument augmentation annotation.

Additionally, annotators were asked to add any event arguments considered valid in Rich ERE, but missed during Rich ERE annotation. For example, place arguments had occasionally not been annotated when they did not occur in the same sentence as the event trigger and other event arguments, and these were added during event argument augmentation. Figure 2 shows a comparison of the increase in number of event arguments through augmentation in 2016 and 2017. In 2016, augmentation increased the number of event arguments (compared to ERE without augmentation) by only 6-7%, but in 2017 there was a 42% increase in Chinese, a 53% increase in English and a 61% increase in Spanish, meaning that many more valid event arguments were captured in the 2017 annotation under the manual augmentation process, as compared with the automated method utilized in 2016.

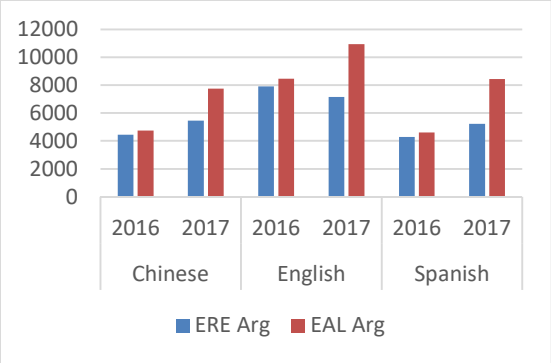

Figure 1: Event Argument Augmentation in 2016 and 2017

3.4 Cross-document Event Coreference

In order to provide additional training data for the 2017 KBP Cold Start evaluation, which required systems to build a corpus-wide KB from scratch using a pre-defined KB schema and a collection of unstructured text, we expanded ERE to include cross-document and cross-lingual event coreference, utilizing the event hopper framework. The goal was to provide a gold standard labeling of which arguments participated in which events across the corpus.

The cross-document and cross-lingual event coreference annotation took existing within-document Rich ERE event hopper annotation as input, and coreferenced event hoppers from different documents. The criteria for judging whether hoppers were coreferential or not were the same as those outlined in the description of within-document event hoppers in Song et al. (2015). Procedurally, an annotator compared an existing event hopper from one document in Rich ERE to an event hopper in another document and decided whether the two event hoppers were coreferential. We used the 505 core source documents which had already been annotated

with Rich ERE for the TAC KBP 2016 evaluations (Ellis et al., 2016).

There were altogether 55,471 hopper pairs judged, and the annotation effort resulted in 892 coreference pairs, with a coreference ratio of 1.6%, as shown in Table 3. The cross-document event hoppers that were judged as coreferential were then clustered as event hopper clusters. This resulted in a total of 389 cross-document event hopper clusters.

Lang	Total pairs	Corefe-rential pairs	Total hoppers	Hopper clusters
Chinese	14,473	256	1,643	108
English	33,520	423	2,454	195
Spanish	7,478	213	1,234	86
Total	55,471	892	5,329	389

Table 4: Cross-document Event Coreference Annotation Results

4 Belief and Sentiment

2017 is the second year for the Belief and Sentiment (BeSt) track for TAC KBP. The goal of the BeSt task is to allow the detection of beliefs and sentiment to augment the information about entities, relations, and events in the knowledge base. To support this goal, belief and sentiment are annotated with respect to entities, relations, and events as annotated in the core set of KBP documents annotated with Rich ERE. BeSt annotation includes labeling the holder of all beliefs and sentiments directed toward taget entities, relations and events from the ERE annotation. The BeSt task for 2017 had one minor change from 2016, which was the addition of "author" or "other" as fillers for the source of a belief or sentiment when the source was not annotated as an ERE entity.

4.1 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. In the definitions below the term "proposition" refers to the existence of the target relation or event and/or the role of entities as event arguments.

Committed Belief (CB) -- the holder believes the proposition with certainty

Non-committed Belief (NCB) -- the holder believes the proposition to be possibly, but not necessarily, true

Reported Belief (ROB) -- the holder reports the belief as belonging to someone else, without specifying his/her own belief or lack of belief

Not Applicable (NA) -- the holder expresses some cognitive attitude other than belief toward the proposition, such as desire, intention, or obligation.

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. 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, in the sentence "ISIS may have been responsible for the bombing," the writer expresses a committed belief that the Conflict.Attack event ("bombing") 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, but they do appear in both the training and gold standard evaluation annotation.

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 proposition is true, while negative polarity means belief that it is not true. Table 6 summarizes the interaction of polarity and committed/non-committed belief for each target type (events, relations, entities).

		Committed	**Non-Committed**
Positive polarity	**Event**	Definitely occurred	Possibly/likely occurred
	Relation	Definitely true	Possibly/likely true
	Entity	Definitely participated in annotated role	Possibly/likely participated in annotated role
Negative polarity	**Event**	Definitely did not occur	Possibly/likely did not occur
	Relation	Definitely false	Possibly/likely false
	Entity	Definitely did not participate in annotated role	Possibly/likely did not participate in annotated role

Table 5: Interpreting Committed Belief and Polarity

Only entity, relation, and event mentions annotated in DEFT Rich ERE can be targets of belief and sentiment annotation. Beliefs and sentiments toward other targets are not annotated. The holders of beliefs and sentiments are entity mentions annotated in Rich ERE, or, when the holder of the belief or sentiment is not annotated as an entity, "author" (for the author of the document) or "other" (for any other source that is not annotated as an ERE entity).

Once the first-pass annotator has completed annotation of both sentiment and belief on a document, a senior annotator reviewed the annotations in a second pass, with a particular focus on sentiment, since lower consistency for sentiment was identified during previous annotation efforts on this task.

4.2 Results

For the 2017 BeSt evaluation, LDC produced gold standard annotation for the evaluation set. No new training data was produced, but the 2016 evaluation data was updated to include the "author" and "other" flags for sources that were previously unannotated. The evaluation data was the core set of ERE annotated documents used in other KBP tracks. The table below provides information about the quantities and distributions of annotations in the 2016 and 2017 evaluation data across the three languages. Note that for sentiment, the annotation counts include the value "none" for sentiment, since annotators consider each potential sentiment target and mark any target that has neither positive nor negative sentiment as "none". The number reported here therefore indicates the number of annotation decisions made. The BeSt evaluation only considers the annotations that are marked as either positive or negative.

Language	Belief annotations		Sentiment annotations	
	2016	2017	2016	2017
Chinese	12,163	18,854	18,982	23,761
English	21,188	20,030	25,358	22,370
Spanish	12,546	15,528	17,353	19,622

Table 6: Total belief and sentiment annotations by language

Differences in total number of belief and sentiment annotations across languages is a result of each language having a slightly different density of ERE annotations (and therefore belief and sentiment targets). In 2016, English had a significantly higher total number of annotations, but in 2017 the difference across languages was smaller.

For all three languages, the distribution of belief types is similar between the 2016 and 2017 evaluation data. Both English and Spanish had slightly more Committed Beliefs and slightly fewer Reported Beliefs in 2017 compared to 2016. Overall, the pattern of very large numbers of Committed Beliefs and very small numbers of Non-Committed Beliefs remains constant across languages as well as between 2016 and 2017, as can be seen in Figure 2.

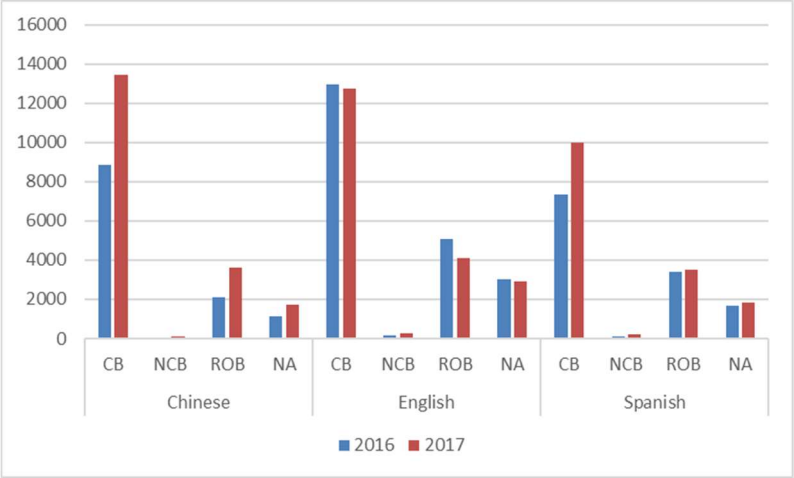

Figure 2: Belief annotations in 2016 and 2017

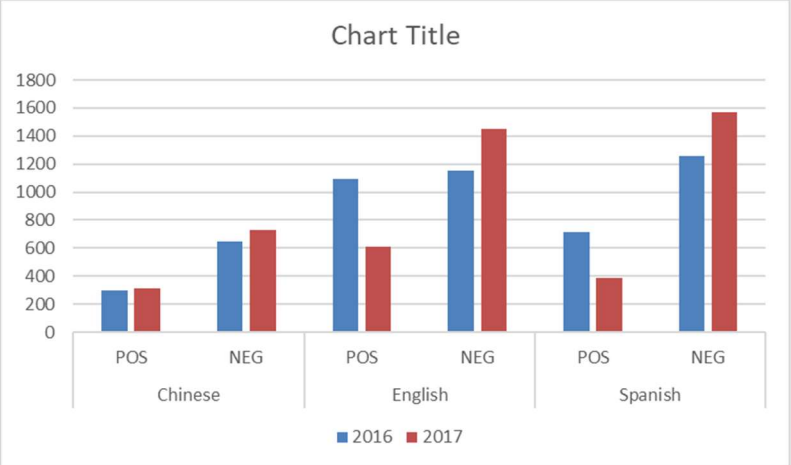

Figure 3: Sentiment annotations in 2016 and 2017

The presence of sentiment in the data remains low from 2016 to 2017 (see Figure 3), and the pattern in which negative sentiment is more frequent than positive sentiment also holds true, across both years and languages, with the 2017 data for English falling closer to the pattern for Chinese and Spanish than in 2017.

5 Cold Start

At a very high level, data development in support of Cold Start for 2017 was relatively consistent with the approach used in 2016. 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 changes to Cold Start data development were necessary in 2017 to support the addition of sentiment slots, which sought to extract positive and negative sentiment held by entities towards other entities, as well as the addition of a new set of slots, based on the event types annotated in the event tracks, that sought to extract the events in which entities in the source corpus are somehow involved.

Unlike all the other tracks discussed up to this point, Cold Start is the only TAC KBP 2017 data that did not directly utilize ERE data as input. Since queries and responses for the Cold Start manual run were to come from across the full 90K corpus, there were less advantages to be had from using the ERE data, as compared with other KBP tracks, since ERE was restricted to the manually-selected 500-document subset.

5.1 Query and Manual Run Development

The most consequential changes to query development in 2017 were made in order to support the production of queries utilizing more than one category of slot (relation, sentiment, and/or event). 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 0-hop query, 2 slots for a 1-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 annotator (second pass) who exhaustively annotated all responses they could find for the queries developed in the first pass. Starting in 2016, kits for multilingual queries, however, required review by up to 6 annotators – one for each language in the first pass, and one for each language in the second pass. With the addition of event and sentiment slots, kits required as many as 12 passes – nine initial query development passes (one for each of the three slot categories, in each of the three languages), as well as an exhaustive second pass in each language.

Total queries	1,392
Total entry-point entities	237
Total manual run responses	3,495

Table 7: 2017 Cold Start data volumes

5.2 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.

Unlike query and manual run development, assessment was not significantly affected by the addition of sentiment and event slots. This is because, unlike query development, which requires annotators to work with slots of multiple categories simultaneously, each assessment kit deals only with one entity and one slot at a time, meaning that, at most, each kit needs 3 passes (one for each language). It's worth noting, however, that the addition of sentiment and event slots meant that assessors had to be familiar with the definitions of 78 different KBP slots in 2017, as opposed to 41 in previous years.

5.3 Results

Results were positive for LDC's manual run in 2017 as compared to the previous year. Cross-lingual precision and recall improved, as did monolingual precision for each language. Monolingual English and Spanish recall also improved; only Chinese recall remained static. We believe the Chinese recall was affected, at least in part, by a handful of queries that proved to be highly productive in Chinese for systems. Note that this paper reports preliminary 2017 results that are available as of submission.

Year	Lang	Precision	Recall	F1
2017	ENG	94%	41%	58%
2017	CMN	88%	25%	40%
2017	SPA	88%	78%	83%
2017	X-ling	90%	38%	54%
2016	ENG	80%	34%	48%
2016	CMN	76%	25%	38%
2016	SPA	87%	64%	74%
2016	X-ling	78%	35%	49%

Table 8: LDC's manual run scores for Cold Start

6 Conclusion

This paper discussed the linguistic resources produced in support of the TAC KBP 2017 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 further analysis of 2017 results, 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.

7 References

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, Stephanie M. Strassel. 2016. Overview of Linguistic Resources for the TAC KBP 2016 Evaluations: Methodologies and Results. *TAC KBP Workshop 2016*: National Institute of Standards and Technology, Gaithersburg, MD, November 14-15.

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: LDC Data Distributed to TAC KBP 2017 Participants

Catalog ID	Title	Release Date
LDC2014T16	TAC KBP Reference Knowledge Base	*all pre-2017 data*
LDC2015E17	TAC KBP Chinese Entity Linking Comprehensive Training and Evaluation Data 2011 - 2014	*all pre-2017 data*
LDC2015E19	TAC KBP English Entity Linking Comprehensive Training and Evaluation Data 2009 - 2013	*all pre-2017 data*
LDC2015E42	TAC KBP Knowledge Base II - BaseKB	*all pre-2017 data*
LDC2015E45	TAC KBP Comprehensive English Source Corpora 2009-2014	*all pre-2017 data*
LDC2015E46	TAC KBP English Regular Slot Filling Comprehensive Training and Evaluation Data 2009-2014	*all pre-2017 data*
LDC2015E47	TAC KBP English Sentiment Slot Filling Comprehensive Training and Evaluation Data 2013-2014	*all pre-2017 data*
LDC2015E49	TAC KBP English Surprise Slot Filling Comprehensive Training and Evaluation Data 2010	*all pre-2017 data*
LDC2015E50	TAC KBP English Temporal Slot Filling Collected Training and Evaluation Data Sets 2011 and 2013	*all pre-2017 data*
LDC2016T26	TAC KBP Spanish Entity Linking Comprehensive Training and Evaluation Data 2012 - 2014	*all pre-2017 data*
LDC2016E35	TAC KBP Chinese Regular Slot Filling Comprehensive Training and Evaluation Data 2014	*all pre-2017 data*
LDC2016E63	TAC KBP 2016 Evaluation Source Corpus V1.1	*all pre-2017 data*
LDC2016E114	TAC KBP 2016 Belief and Sentiment Evaluation Gold Standard Annotation	*all pre-2017 data*
LDC2017E02	TAC KBP Event Nugget Detection and Coreference Comprehensive Training and Evaluation Data 2014-2016	*all pre-2017 data*
LDC2017E03	TAC KBP Entity Discovery and Linking Comprehensive Training and Evaluation Data 2014-2016 V1.1	*all pre-2017 data*
LDC2017E04	TAC KBP Cold Start Comprehensive Evaluation Data 2012-2016	*all pre-2017 data*
LDC2017E05	TAC KBP Event Argument Comprehensive Training and Evaluation Data 2014 - 2016	*all pre-2017 data*

Table 1: Prior TAC KBP Data Sets Distributed in 2017 for System Training and Development

Track	Catalog ID	Title
All	LDC2017E25	TAC KBP 2017 Evaluation Source Corpus V1.1
Cold Start	LDC2017E26	TAC KBP 2017 Cold Start Evaluation Queries V1.1
Cold Start	LDC2017E34	TAC KBP 2017 Cold Start Evaluation Queries and Manual Run V1.2
All	LDC2017E51	TAC KBP 2017 Evaluation Core Source Corpus
ED&L	LDC2017E52	TAC KBP 2017 Entity Discovery and Linking Evaluation Gold Standard Entity Mentions and Knowledge Base Links
BeSt	LDC2017E53	TAC KBP 2017 Eval Core Set Rich ERE Annotation
Event Nugget	LDC2017E54	TAC KBP 2017 Eval Core Set Event Nugget Annotation
Event Argument	LDC2017E55	TAC KBP 2017 Eval Core Set Rich ERE Annotation with Augmented Event Arguments
Cold Start	LDC2017E56	TAC KBP 2017 Cold Start Evaluation Assessment Results V2.0
BeSt	LDC2017E80	TAC KBP 2017 Belief and Sentiment Evaluation Gold Standard Annotation

Table 2: Newly Created TAC KBP Data Sets