Laying the Groundwork for Knowledge Base Population: Nine Years of Linguistic Resources for TAC KBP

Jeremy Getman, Joe Ellis, Stephanie Strassel, Zhiyi Song, Jennifer Tracey
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104
{jgetman, joellis, strassel, zhiyi, garjen}@ldc.upenn.edu

Abstract

Knowledge Base Population (KBP) is an evaluation series within the Text Analysis Conference (TAC) evaluation campaign conducted by the National Institute of Standards and Technology (NIST). Over the past nine years TAC KBP evaluations have targeted information extraction technologies for the population of knowledge bases comprised of entities, relations, and events. Linguistic Data Consortium (LDC) has supported TAC KBP since 2009, developing, maintaining, and distributing linguistic resources in three languages for seven distinct evaluation tracks. This paper describes LDC's resource creation efforts for the various KBP tracks, and highlights changes made over the years to support evolving evaluation requirements.

Keywords: knowledge base population, information extraction, linguistic resources

1. Introduction

Text Analysis Conference (TAC) an annual series of open technology evaluations organized by the National Institute of Standards and Technology (NIST). The Knowledge Base Population (KBP) evaluation track (McNamee et al. 2010) encourages the development of systems that can extract information from unstructured multilingual text and in order to populate an existing or emergent knowledge base. Since the start of TAC KBP in 2009 Linguistic Data Consortium (LDC) has provided linguistic resources by building labeled training and test sets and by assessing system results.

Each year's KBP evaluation comprises a number of component evaluation tasks. Over time the KBP evaluations, and by extension the data to support those evaluations, has moved in the direction of greater complexity, with more integration of individual evaluation tracks and data sets, a greater emphasis on multilingual data for all tracks, and an increasing focus on events (in addition to entities and slots). These developments culminated in 2017 with an expanded end-to-end Cold Start track that evaluated systems' ability to combine the other KBP tracks and extract all entities, relations, sentiments, and events from a document collection, perform corpus-wide clustering of the extracted items, and build a structured knowledge base from scratch (NIST, 2017).

In the sections that follow we describe the KBP corpora developed by LDC for each of the seven primary evaluation tracks, including source data, knowledge bases and annotation and assessment efforts between 2009-2017. TAC KBP will continue in 2018 with new evaluation tracks.

2. Source Data and Knowledge Bases

Source data for the KBP evaluations consists of English, Spanish and Chinese text in two primary genres: formal newswire (NW) and informal web text (primarily blogs and discussion forums) drawn from existing LDC collections as well as newly collected material. In 2009-2015 we selected a separate set of documents for each evaluation track. Starting in 2016, the consolidation of evaluations into a smaller number of more complex and challenging tasks led to the use of a single set of test data shared by all tasks. This put additional demands on data selection given the need for a single corpus to support multiple, sometimes mutually contradictory, requirements. For instance, the list of required features for the 2016 source data included all of the following:

- Roughly equal representation of all three languages (ENG, SPA, CMN)
- Roughly equal representation of both genres (formal and informal)
- Corpus should cover a relatively short time span
- 800 tokens max per document excluding quote regions for all 90K documents
- Discussion forum documents will start from the beginning of the thread
- Each event sub/type in the ontology must be present in each genre/language
- Each event type and subtype in the ontology must have at least one mention in each of 30 or more documents in each language
- Cross-lingual event hoppers in the corpus, for at least half of the event types in the ontology, made up of 2-3 event mentions
- 50 instances of relatively simple, non-confusable events that are mentioned in 3 or more documents
- 10 or more event hoppers with mentions in 10 or more documents
- Presence of only unnamed mentions for some specific, individual entities in some documents
- Presence of synonymous entities (entities referred to by more than one non-matching string in the corpus)
- Presence of polysemous entities (distinct entities referred to by equivalent strings in the corpus)

Table 1 below summarizes KBP source data provided by LDC for each year's evaluations.

For every new evaluation track introduced to KBP, LDC produced a small amount of labeled training data to illustrate the annotation approach and data properties for the track. After this initial training set no new labeled training data was produced; instead, labeled test sets from prior years were released to evaluation participants for use as training and development data. LDC also made related

resources produced under other efforts available to KBP participants to use as supplemental training data.

Year	Genre	ENG	CMN	SPA
2009	formal	1,286,609	-	-
	informal	3,040	-	-
2010	formal	1,287,292	-	-
	informal	490,596	-	-
2011	formal	1,287,292	1,000,000	-
	informal	490,596	-	-
2012	formal	2,287,549	2,000,256	1,000,020
	informal	1,490,595	815,886	-
2013	formal	1,000,257	2,000,256	910,734
	informal	1,099,062	1,015,027	-
2014	formal	1,000,562	2,000,256	910,734
	informal	1,099,423	1,015,027	649,095
2015	formal	9,270	84	84
	informal	40,521	82	83
2016	formal	15,000	15,000	15,000
	informal	15,000	15,000	15,000
2017	formal	15,000	15,000	15,000
	informal	15,000	15,000	15,000

Table 1: TAC KBP evaluation source documents

In addition to creating the training and test corpora required for each year's evaluation tracks, LDC also produced the knowledge bases used for evaluation. KBP has used two distinct KBs since its inception. Prior to the first KBP evaluation in 2009, LDC created a KB based on a Wikipedia snapshot, extracting information from page titles, infoboxes and article text from over 800,000 entries (Simpson et. al., 2010). Each node in the reference KB corresponds to a Wikipedia page for a person, organization, or geopolitical entity and consists of predefined attributes derived from infoboxes (Linguistic Data Consortium, 2014). In 2015 a decision was made by evaluation coordinators to instead use BaseKB (http://basekb.com), which is a subset of Freebase represented in RDF (Linguistic Data Consortium, 2015). Because of the additional complexity of BaseKB compared to the Wikipedia-based KB, LDC prepared a custom human-readable version for use during annotation and assessment.

3. Evolution of Data for KBP Evaluation Tracks: 2009-2017

3.1 Entity Linking

Entity Linking (EL) requires systems to link named mentions of person, organization, and geopolitical entities in text to entries in a knowledge base (KB), report if no matching entries exist, and group mentions without entries according to identity coreference. Entity Linking began in 2009 in English (Simpson et al., 2010), added a Chinese cross-lingual version of the task in 2011, and further expanded with a Spanish version of the task in 2012 (Ellis et al., 2012; Li et al., 2012). The 2013 EL task remained largely unchanged from the tri-lingual version established in 2012. In 2014, only cross-lingual Chinese and Spanish EL evaluations were held, as English was replaced by Entity Discovery & Linking.

Although the languages in Entity Linking changed over the track's six years, the goals of query selection for Entity Linking did not (Ellis et al., 2014). Annotators sought to collect the most confusable named entity mentions they could find in the corpus. A query's confusability was measured both by the number of distinct entities in the set of queries that are referred to by its namestring (polysemy) as well as the number of distinct namestrings in the pool that refer to the entity (synonymy). Entity Linking queries were selected with the intention of representing as evenly as possible the three entity types, as well as the ratio of entities in the KB to those not. For cross-lingual EL, English entity mentions co-referential with non-English queries were selected when possible.

The following table summarizes Entity Linking resources developed by LDC.

Year	Source Documents			Queries		
	ENG	CMN	SPA	ENG	CMN	SPA
2009	3688	-	-	3904	-	-
2010	3684	-	-	3750	-	-
2011	2231	4329	-	2250	4347	-
2012	2016	2271	3772	2226	2280	3890
2013	1820	2143	1832	2190	2155	2117
2014	-	2860	2207	-	3253	2596

Table 2: Entity Linking resources

3.2 Entity Discovery & Linking

A new variant of Entity Linking, named Entity Discovery & Linking (EDL), was performed for the first time in 2014 in English. The goal of EDL is full entity extraction from a collection of documents, followed by linking entities to a KB and clustering any entities not in the KB. EDL differed from Entity Linking in that systems and annotators exhaustively annotated documents to create a gold standard for system scoring, instead of cherry-picking ambiguous entities from the corpus (Ellis et al., 2014). While the 2014 task dealt only with named mentions of person, organization, and geopolitical entities, locations and facilities were also annotated in subsequent years (2015-2017), as were nominal mentions. EDL was expanded from English-only to English, Chinese, and Spanish starting in 2015.

In 2014-2015, EDL gold standard annotation required identification and classification of all valid mentions of the targeted entity types within the source corpus (Ellis et al., 2015). Titles were also annotated in 2015 to help systems distinguish between titles and nominal mentions of persons (e.g. "president"). In 2016-2017, rather than starting with a blank slate, annotators worked with entity mentions from Entities, Relations, and Events (ERE), an annotation task developed by LDC for DARPA's Deep Exploration and Filtering of Text program (DEFT) (DARPA, 2012). ERE exhaustively labels entities, relations and events, along with their attributes, according to specified taxonomies (Song et al., 2015). In EDL, ERE entity annotations were displayed in the context of their source documents, so annotators could check for errors and misses, as well as ERE annotations at variance with EDL guidelines (though correct for ERE). In all years,

labeled entities were then linked to the KB or marked as NIL (not in the KB) or Unknown (insufficient information in the source data to know if an entity was in the KB). After document-level KB linking, senior annotators then performed cross-document, cross-language NIL clustering, aided by English descriptions of non-English entities.

The following table summarizes Entity Discovery & Linking resources developed by LDC.

Year	Task	Source Documents	Queries		
			ENG	CMN	SPA
2014	training	138	5598	-	-
2014	eval	160	6349	-	-
2015	pilot	15	200	266	220
2015	training	444	13545	13116	4177
2015	eval	500	15645	11066	5822
2016	eval	505	9231	8845	6964
2017	eval	500	6915	10246	7212

Table 3: Entity Discovery & Linking resources

3.3 Slot Filling

The regular Slot Filling (SF) task involves mining information about entities from text. Systems and LDC annotators search a corpus for information about persons and organizations and add new information to an existing knowledge base. LDC produced data in support of English Slot Filling from 2009 through 2014. In 2014, LDC also produced data for a Chinese Slot Filling pilot evaluation. In addition to regular Slot Filling, LDC also produced data in support of Surprise Slot Filling, in 2010, Temporal Slot Filling, in 2011 and 2013 (Ellis et al., 2013), and Sentiment Slot Filling, in 2013 and 2014. The Surprise and Sentiment SF tasks followed the same guidelines as regular SF with the exception that the slots evaluated targeted novel types of information. Surprise SF added four slots in the same vein as regular SF slots, and Sentiment SF sought to extract positive and negative sentiment held by entities toward other entities, including geopolitical entities (unlike regular SF). Temporal SF utilized the regular SF slots, and involved annotation of temporal information indicating when a given SF relation held true.

Entities (the basis of SF queries), were selected for their non-confusability and productivity. An entity was considered non-confusable if there existed one or more canonical references to it in the source corpus. Productivity was determined by searching the source corpus to find whether at least two answers existed for the entity. After 2009, LDC also developed manual runs, the set of valid human-produced responses to each of the SF queries. From 2012, responses included a justification (the minimum text extract from the source corpus supporting the validity of a response). Valid justifications included all three elements of a relation: its subject entity, slot type, and answer. During assessment, annotators judged the validity of human and system responses, and grouped instances of the same response. Answers were marked correct if they adhered to the slot definitions and were supported in the text. For the 2009-2013 evaluations, attributes in the KB were mapped to the set of SF queries before assessment, thereby indicating returned responses' redundancy with the KB.

The following tables summarize Slot Filling resources developed by LDC.

Year	Task	Lang.	Queries	LDC Responses	Assessed Responses
2009	eval	ENG	53	-	10416
2010	training	ENG	98	336	-
2010	eval	ENG	100	799	24515
2011	training	ENG	198	1627	-
2011	eval	ENG	100	796	28041
2012	eval	ENG	80	1553	22885
2013	eval	ENG	100	2383	27655
2014	eval	ENG	100	2216	21956
2014	training	CMN	32	967	-
2014	eval	CMN	103	2858	2878

Table 4: Regular Slot Filling resources

Year	Task	Lang.	Queries	LDC Responses	Assessed Responses
2010	training	ENG	32	83	-
2010	eval	ENG	40	252	996

Table 5: Surprise Slot Filling resources

Year	Task	Lang.	Queries	LDC Responses	Assessed Responses
2013	training	ENG	163	986	-
2013	eval	ENG	160	977	5160
2013	dual	ENG	-	-	1145
2014	eval	ENG	400	594	6383

Table 6: Sentiment Slot Filling resources

Year	Task	Lang.	Queries	LDC Responses	Assessed Responses
2011	training	ENG	50	1258	-
2011	eval	ENG	100	1413	-
2013	training	ENG	7	16	-
2013	eval	ENG	271	1519	2035

Table 7: Temporal Slot Filling resources

3.4 Event Argument Linking

The Event Argument Linking (EAL) task requires systems and annotators to extract event arguments (entities or attributes playing a role in an event), indicate their role, link the arguments involved in the same event, and format the information in a manner suitable as input to a knowledge base. LDC produced data in support of EAL in each of its four years, from 2014-2017. In 2014 and 2015, the EAL evaluation used an assessment paradigm; 2016 and 2017 instead used a gold standard. In the 2014 EAL evaluation, annotators marked one mention of each valid, unique event argument within the EAL source corpus. In 2015, the task was expanded to include the linking of related event arguments; annotators marked each unique

event argument and clustered related arguments into event hoppers.

For 2014 assessment, annotators performed coreference on all responses to a document, then judged the parts of each response – the event type, the role a response played in the event, and the entity filling that role. Annotators also indicated each response's modality as well as its mention type (name or nominal). In 2015, a final step was added, wherein annotators grouped responses into event hoppers (Song et al., 2015) to indicate a type of event coreference. In 2016, instead of an assessment paradigm, LDC created a set of gold standard EAL annotations against which system submissions were scored (Ellis et al., 2016). The gold standard was an expanded version of ERE data augmented by a script developed by the EAL evaluation track coordinators. Annotators reviewed the results of this augmentation, which added inferred arguments invalid for ERE (but not EAL) and/or difficult for annotators to find. The same approach was taken in 2017, except that instead of reviewing automatically generated augmentations, annotators performed a fully manual augmentation pass, which increased the number of augmented event arguments, as compared with 2016. An English cross-document task was also added in 2016 only, in which LDC selected queries, produced responses, and assessed human and system responses. Annotators selected queries comprised of single event arguments each indicating an event hopper in the EAL gold standard, and searched for all responses to those queries. During assessment, annotators decided if a response's justification proved that a document contained an instance of the corresponding query event.

The following table summarizes Event Argument Linking resources developed by LDC.

Year	Task	Lang.	Source Docs	LDC Responses	Assessed Responses
2014	pilot	ENG	60	-	32054
2014	eval	ENG	528	5947	57599
2015	training	ENG	55	-	-
2015	eval	ENG	500	5207	45391
2016	gold standard eval	ENG, CMN, SPA	505	17809	-
2016	x-doc pilot	ENG	2092	98	2689
2016	x-doc eval	ENG	30000	700	7697
2017	gold standard eval	ENG, CMN, SPA	500	27109	-

Table 8: Event Argument Linking resources

3.5 Event Nugget

The Event Nugget track seeks to evaluate system performance in detection and coreference of event references in text (Mitamura et al., 2015). 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 (Ellis et al., 2015). Event Nugget started as a pilot evaluation within the DEFT program in 2014. In 2015, event nuggets were redefined to align with the treatment of events in DEFT Rich ERE (Song et al., 2015). Also in 2015, coreference of event nuggets was added, using the definition of event hoppers developed in Rich ERE. In 2016 and 2017, there was no separate annotation task conducted solely to support the Event Nugget evaluations; the data were entirely produced by running a script over ERE data to extract and reformat a subset for use by Event Nugget. Additionally, Chinese, Spanish, and English source documents were used as inputs in 2016 and 2017, whereas the task had been English-only in previous iterations.

Year	Task	Lang.	Source Docs	Event Nuggets	Event Hoppers
2014	training	ENG	151	3782	-
2014	eval	ENG	200	6921	-
2015	training	ENG	446	12301	7481
2015	eval	ENG	202	6438	4125
2016	eval	ENG, CMN, SPA	500	9042	6799
2017	eval	ENG, CMN, SPA	500	11687	8039

Table 9: Event Nugget resources

3.6 Belief and Sentiment

Belief and Sentiment (BeSt), part of TAC KBP in 2016 and 2017, 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 (Ellis et al., 2016). BeSt requires that belief and sentiment be annotated with respect to entities, relations, and events as annotated in ERE. Entities can be holders/reporters of belief and sentiment, as well as targets of sentiment; relations and events can be the targets of belief and/or sentiment. BeSt annotation also labels an entity's role in an event as a target of belief, separate from belief in the event itself. 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. Belief annotation marks the belief-holder's commitment to a belief in the occurrence of an event (event-target), the participation of an entity in an annotated event (entity-target), and/or the existence of a relation (relation-target). In addition to the target and belief-type, the holder of the belief is explicitly indicated, as is the polarity of the belief. Positive and negative sentiment is annotated with entities, relations, and events as targets, and, as in Belief annotation, the holder of the sentiment is indicated.

Table 10 below summarizes BeSt resources developed by LDC.

Year	Task	Lang.	Source Docs	Belief Annotations	Sentiment Annotations
2016	eval	ENG, CMN, SPA	494	45897	61693
2017	eval	ENG, CMN, SPA	500	54412	65753

Table 10: Belief and Sentiment resources

3.7 Cold Start

Cold Start, part of TAC KBP from 2012 through 2017, is designed to evaluate a system's ability to construct a new knowledge base from the information provided in a text collection, by combining technologies developed via other KBP tracks. Like Slot Filling, Cold Start involves mining information about entities from text, and as in Entity Discovery & Linking, Cold Start systems must also find all entities mentioned in the text (Ji et al., 2016). From 2012 through 2016, Cold Start focused on person, organization, and geopolitical entities; facilities and locations were added in 2017. From 2012 to 2015, Cold Start was English-only, but following a Spanish-English pilot in 2016 the track expanded to English, Chinese, and Spanish for 2016 and 2017. On top of regular Slot Filling slots, the focus of Cold Start since 2012, the 2017 task also incorporated the Sentiment SF slots, as well as a new set of event-focused slots, derived from Event Argument Linking, that sought to extract the events in which entities were involved (Getman et al., 2017).

In Cold Start query development, annotators created queries defined by an entity initiating a chain of relations. Unlike Slot Filling, which generates single binary relations, Cold Start strings multiple relations together, so that the object of one relation becomes the subject of another. An example query could be phrased: "Find all shareholders of organizations at which Jane Doe has been an employee." After developing queries, annotators searched for all responses to those queries that could be found in the source data. Responses included justification, extents of text proving the validity of a response. Through 2015, Cold Start queries and responses were developed concurrently, such that annotators could switch between investigating candidate queries and annotating responses. In 2016 and 2017, however, query and response development were separated, as a result of the addition of Spanish and Chinese, which meant a single annotator could no longer annotate all responses to a query, as each query required searching for answers in three languages. In assessment, annotators judged human and system responses. Cold Start followed the same paradigm as that in Slot Filling assessment. However, unlike Slot Filling, because Cold Start queries involved chains of slots, Cold Start assessment necessarily happened in multiple stages. The first stage mirrored Slot Filling assessment, but the second stage involved assessment of slots branching off of answers marked correct in the first stage.

The following table summarizes Cold Start resources developed by LDC.

Year	Task	Lang.	Queries	LDC Responses	Assessed Responses
2012	eval	ENG	385	979	5015
2013	eval	ENG	326	1627	6745
2014	eval	ENG	247	1386	7254
2015	eval	ENG	2539	2218	30654
2016	pilot	SPA, ENG	2118	1238	818
2016	eval	ENG, CMN, SPA	1077	4739	25416
2017	eval	ENG, CMN, SPA	1392	3495	26802

Table 11: Cold Start resources

4. Discussion and Conclusions

We have described LDC's resource creation efforts in support of TAC KBP evaluations since 2009, highlighting changes in our approach required to meet new requirements within and across KBP tasks. In considering the evolution of KBP and its data over time, one strong trend has been the consolidation of simpler monolingual tasks (like English Entity Linking into more challenging multilingual tasks (like Trilingual Entity Discovery and Linking and Cold Start). In part this evolution reflects the maturation of KBP technology, but it also highlights the fact that starting in 2012 KBP also served as the primary framework for evaluating system capabilities in the DARPA Deep Exploration and Filtering of Text (DEFT) program (DARPA, 2012). The goal of the DEFT program is to develop technologies capable of extracting knowledge from unstructured text in multiple languages and genres. DEFT's growing focus on multi-lingual technologies and whole-corpus (as opposed to sentence- or document-level) understanding is reflected in the evolution of KBP tracks over the past 5 years. The table below presents a summary of KBP tracks and their languages between 2009-2017.

	2009	2010	2011	2012	2013	2014	2015	2016	2017
Entity Linking	E	E	E,C	E,C,S	E,C,S	C,S	n/a	n/a	n/a
Entity Discovery & Linking	n/a	n/a	n/a	n/a	n/a	E	E,C,S	E,C,S	E,C,S
Regular Slot Filling	E	E	E	E	E	E,C	n/a	n/a	n/a
Surprise Slot Filling	n/a	E	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Sentiment Slot Filling	n/a	n/a	n/a	n/a	E	E	n/a	n/a	n/a
Temporal Slot Filling	n/a	n/a	E	n/a	E	n/a	n/a	n/a	n/a
Event Argument Linking	n/a	n/a	n/a	n/a	n/a	E	E	E,C,S	E,C,S
Event Nugget	n/a	n/a	n/a	n/a	n/a	E	E	E,C,S	E,C,S
BeSt	n/a	n/a	n/a	n/a	n/a	n/a	n/a	E,C,S	E,C,S
Cold Start	n/a	n/a	n/a	E	E	E	E	E,C,S	E,C,S

Table 12: Increasing complexity and multilinguality in KBP over time

As illustrated here, KBP data introduced greater complexity over time, with more integration of distinct evaluation tracks and data sets, a greater emphasis on multilingual data for all tracks, and an increasing focus on events (as well as entities and slots). In 2013 we expanded Slot Filling to include data both Sentiment Slot Filling and Temporal Slot Filling. In 2014 we added data for two

event-focused tracks Event Argument Linking and Event Nugget, and expanded Entity Linking annotation to cross-document entity extraction and clustering for English (EDL); Chinese and Spanish EDL followed in 2015, which meant cross-document clustering was necessarily also cross-lingual. In 2016 we made the move to using the same source corpus for all KBP evaluation tracks, and we expanded Cold Start from monolingual English to cross-lingual English/Chinese/Spanish; we also added new information about entities, relations, and events with beliefs and sentiment. In 2017, Cold Start was further expanded to include event and sentiment slots, making Cold Start very nearly the sum total of all component KBP evaluations, testing extraction and clustering of entities, relations, events, and sentiment. In all LDC has produced over 150 distinct KBP corpora, comprising over to 150,000 queries, 84,000 manual runs and 310,000 system assessments.

After the conclusion of an evaluation series, resources are consolidated into one or more comprehensive packages and released into LDC's public catalog, making them generally available for language-related research, education and technology development. To date LDC has published several KBP corpora including the pre-2015 knowledge base (LDC2014T16); training and evaluation data for cross-lingual entity linking in Spanish (LDC2016T26) and Chinese (LDC2017T17); and the source data used in all English evaluations between 2009-2014 (LDC2018T03). An additional 15-20 KBP corpora will be published in the catalog in the coming months and years. The TAC KBP evaluation series will continue into 2018 with the introduction of new tracks.

5. Bibliographical References

Dang, H.T., Lin, J., and Kelly, D. 2006. Overview of the TREC 2006 Question Answering Track. In Fifteenth Text Retrieval Conference (TREC 2006) Proceedings, Gaithersburg, MD.

DARPA. 2012. *Broad Agency Announcement: Deep Exploration and Filtering of Text (DEFT)*. Defense Advanced Research Projects Agency, DARPA-BAA-12-47.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. 2004. Automatic Content Extraction (ACE) program - task definitions and performance measures. In Proceedings of the Fourth International Language Resources and Evaluation Conference, Lisbon, Portugal.

Ellis, J., Li, X., Griffitt, K., Strassel, S., and Wright, J. 2012. Linguistic Resources for 2012 Knowledge Base Population Evaluations. TAC KBP Workshop 2012: National Institute of Standards and Technology, Gaithersburg, MD.

Ellis, J., Getman, J., Mott, J., Li, X., Griffitt, K., and Strassel, S. 2013. Linguistic Resources for 2013 Knowledge Base Population Evaluations. TAC KBP Workshop 2013: National Institute of Standards and Technology, Gaithersburg, MD.

Ellis, J., Getman, J., and Strassel, S. 2014. Overview of Linguistic Resources for the TAC KBP 2014 Evaluations: Planning, Execution, and Results. TAC KBP Workshop 2014: National Institute of Standards and Technology, Gaithersburg, MD.

Ellis, J., Getman, J., Fore, D., Kuster, N., Song, Z., Bies, A., and Strassel, S. 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.

Ellis, J., Getman, J., Kuster, N., Song, Z., Bies, A., and Strassel, S. 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.

Getman, J., Ellis, J., Song, Z., Tracey, J., and Strassel, S. 2017. Overview of Linguistic Resources for the TAC KBP 2017 Evaluations: Methodologies and Results. TAC KBP 2017 Workshop: National Institute of Standards and Technology, Gaithersburg, MD.

Ji, H., Nothman, J., Dang, H. T., & Hub, S. I. (2016). Overview of TAC-KBP2016 Tri-lingual EDL and its impact on end-to-end Cold-Start KBP. Proceedings of TAC 2016.

Li, X., Strassel, S. M., Ji, H., Griffitt, K., & Ellis, J. (2012). Linguistic resources for entity linking evaluation: from monolingual to cross-lingual. Annotation. In Proceedings of LREC 2012.

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., & Strassel, S. (2015). Event nugget annotation: Processes and issues. In Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation (pp. 66-76).

McNamee, P., Dang, H.T., Simpson, H., Schone, P., and Strassel, S. (2010). An Evaluation of Technologies for Knowledge Base Population. In Proceedings of the Seventh International Language Resources and Evaluation Conference (LREC), Valleta, Malta.

NIST. (2017). TAC 2017 Cold Start KB Track. https://tac.nist.gov/2017/KBP/ColdStart/index.html. May 2017. Accessed September, 2017.

Simpson, H., Strassel, S., Parker, R., and McNamee, P. 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.

Song, Z., Bies, A., Riese, T., Mott, J., Wright, J., Kulick, S., Ryant, N., Strassel, S., and Ma, X. 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).

Song Z, Bies A, Strassel S, Ellis J, Mitamura T, Dang HT, Yamakawa Y, Holm S. (2016). Event nugget and event coreference annotation. In Proceedings of the The 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation (pp. 37-45).

Walker, C., Strassel, S., Medero, J., and Maeda, K. 2006. ACE 2005 Multilingual Training Corpus. Linguistic Data Consortium, LDC Catalog No.: LDC2006T06.

6. Language Resource References

Simpson, Heather, et al. TAC KBP Reference Knowledge Base LDC2014T16. Web Download. Philadelphia: Linguistic Data Consortium, 2014.

Ellis, Joe, Jeremy Getman, and Stephanie Strassel. TAC KBP Spanish Cross-lingual Entity Linking - Comprehensive Training and Evaluation Data 2012-2014 LDC2016T26. Web Download. Philadelphia: Linguistic Data Consortium, 2016.

Ellis, Joe, Jeremy Getman, and Stephanie Strassel. TAC KBP Chinese Cross-lingual Entity Linking - Comprehensive Training and Evaluation Data 2011-2014 LDC2017T17. Web Download. Philadelphia: Linguistic Data Consortium, 2017.

Ellis, Joe, et al. TAC KBP Comprehensive English Source Corpora 2009-2014 LDC2018T03. Web Download. Philadelphia: Linguistic Data Consortium, 2018.