Overview of Linguistic Resource for the TAC KBP 2014 Evaluations: Planning, Execution, and Results

Joe Ellis, Jeremy Getman, Stephanie Strassel
Linguistic Data Consortium, University of Pennsylvania
{joellis, jgetman, 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 2014, TAC KBP’s sixth year of operation, the evaluations focused on six tracks targeting information extraction and question answering technologies: Entity Linking, Slot Filling, Sentiment Slot Filling, Cold Start, Event Argument Extraction, and Entity Discovery & Linking. Linguistic Data Consortium (LDC) at the University of Pennsylvania has supported the TAC KBP evaluations since 2009 and continued in 2014, maintaining and distributing existing linguistic resources and producing new data, 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 2014.

1 Introduction

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

Since 2009 LDC has been the primary data provider for the evaluation series, developing and distributing training and evaluation datasets as well as tools and specifications. In 2014, LDC created a total of 52 new data sets in support of the KBP evaluations, 6 more than were produced for the 2013 evaluations, and 31 of which were primarily developed between July and September. In addition to official training and evaluation data releases for performers, these packages included preliminary releases for data previews, updates to improve quality and add new data to existing packages, and supplemental releases distributed only to coordinators to support their own data production efforts.

TAC KBP 2014 comprised 6 separate evaluation tracks - Entity Linking, Regular Slot Filling, Cold Start, Sentiment Slot Filling, Entity Discovery & Linking, and Event Argument Extraction - the last two being new tracks for the evaluation series. In order to provide the data needed to support these evaluations, LDC engaged in 18 separate tasks in the months preceding and following the evaluations in July. These tasks can be generally classified as source data selection, query development, annotation (specifically the development of the human-generated 'manual runs' included in the set of assessed responses), or assessment.

This paper describes the full resource creation efforts for TAC KBP 2014. Section 2 describes the planning processes by which the task descriptions and data specifications developed; section 3 gives a brief overview of the source, training, and evaluation data sets made available to KBP performers in 2014 (complete lists of the data are included as appendices at the end of this document); section 4 discusses the procedures and methodologies for data selection, query development, annotation, and assessment for all the 2014 tracks; and section 5 concludes the paper.

2 Planning

Although the TAC KBP 2013 workshop included discussions about the activities for the following year, planning for TAC KBP 2014 did not officially begin until late January, 2014, when coordinators, sponsors, and LDC met to discuss the future evaluations. Over the course of the next four months, coordinators worked to finalize the developing task descriptions, soliciting input from performers regularly. As soon as task descriptions became sufficiently stable, LDC would begin data production, usually starting by updating technical support to meet changes in the tasks and building up teams of annotators to meet volume demands given timelines.

With 2014 marking the 6th year of regular Slot Filling, the track's task description was the first to be completed. Changes to the task included the addition of ambiguous queries that would be shared with other tracks, adding provenance for fillers (in addition to that which was already provided for justification), and marking queries as NIL or non-NIL with respect to live Wikipedia as opposed to the official TAC KBP knowledge base that had been used since 2009 (thus removing from the queries the link to the knowledge base and the resulting list of ‘ignore’ slots – those that could only produce redundant responses).

Status	Modified from 2013
Number of new training releases	0
Number of new evaluation releases	4
Planning begins	January, 2014
Final task description	May, 2014

Table 1: Planning Overview for Slot Filling

Also in its sixth year, English Entity Linking was modified in 2014 to require systems to identify and link (or cluster) all named mentions of entities from provided source documents. The departures from regular Entity Linking were sufficient to consider the task as something completely new - Entity Discovery & Linking - and of particular challenge to LDC in supporting the new task was the desire for far larger datasets than had ever been produced for Entity Linking. As is the case with most new tasks, ED&L remained unstable in the following months, which at one point caused developing training data to require review and updating, following a decision to modify the approach to selecting query name strings to more closely align with that of Entities, Relations, and Events (an annotation task developed for the DEFT program[1] that indicates, for a given document, entities mentioned, the relations between them, and the events in which they participate).

Status	New task
Number of new training releases	1
Number of new evaluation releases	3
Planning begins	January, 2014
Final task description	September, 2014

Table 2: Planning Overview for Entity Discovery & Linking

Spanish and Chinese cross-lingual Entity Linking, in their third and fourth years of operation respectively, remained largely unchanged from 2013. However, two new sets of training data were created for the tasks (one for each language) in order to support queries derived from discussion forum threads[2]. Note that the Entity Linking task description was coupled with that of Entity Discovery & Linking. As such, even though no changes to the task were made as compared to the previous year, the final task description was not released until September.

Status	No changes from 2013
Number of new training releases	2
Number of new evaluation releases	4
Planning begins	January, 2014
Final task description	September, 2014

Table 3: Planning Overview for Entity Linking

Planning for Sentiment Slot Filling came together relatively quickly despite the need to find a new coordinator to lead the track’s development. All parties interested in the task’s development agreed that lowering the barrier of entry for performers should be the primary goal. For 2014, the task was altered such that a single document would serve as the source for responses for each query (as opposed to the whole English corpus, which had been used in 2013). Additionally, as with regular Slot Filling, the 2014 version of Sentiment Slot Filling removed the requirement to link queries or mark them as NIL with respect to an existing knowledge base. This change allowed for discussion forum posters, about whom there is often insufficient information to make a linking decision, to be used as queries.

[1] See the DARPA website for more information about the Deep Exploration and Filtering of Text (DEFT) program - http://www.darpa.mil/Our_Work/I2O/Programs/Deep_Exploration_and_Filtering_of_Text_%28DEFT%29.aspx

[2] Discussion forum threads are harvested online bulletin boards in which individuals (sometimes refered to as ‘posters’) leave and receive responses to messages.

Status	Modified from 2013
Number of new training releases	0
Number of new evaluation releases	3
Planning begins	April, 2014
Final task description	July, 2014

Table 4: Planning Overview for Sentiment Slot Filling

As a new task, the Event Argument Extraction Task sought to surmount some formidable challenges, the most primary of which was formatting information about events to allow for inclusion in a knowledge base. In order to meet the challenges, task coordinators adopted a rigorous planning schedule with early milestones. Training data was developed through the assessment of pilot responses, which were produced manually and by preliminary systems. Assessment was undertaken by both LDC assessors and performers who had submitted responses. In addition to allowing LDC to test its assessment guidelines and technical infrastructure, this process led to more direct involvement by performers in the process of guidelines development.

Status	New task
Number of new training releases	1
Number of new evaluation releases	3
Planning begins	Dec, 2013
Final task description	Sept, 2014

Table 5: Planning Overview for Event Argument Extraction

Cold Start, which was conducted for the third time in 2014, presents the extra complication for planners of having to select a new, hidden source corpus each year. 2014 was no different in this respect but, in an effort to lower the task's barrier of entry, a subset of the readily-processed TAC KBP 2014 English source corpus was used for the task. Changes to the task included the addition of provenance for filler strings, as was done for regular Slot Filling, and an alteration of the output file format for LDC's manual run in order to more closely align with those for regular and Sentiment Slot Filling while still supporting Cold Start's need to link responses to queries via 'hops' (see section 4.2 for further details).

Status	Modified from 2013
Number of new training (pilot) releases	0
Number of new evaluation releases	4
Planning begins	January, 2014
Final task description	August, 2014

Table 6: Planning Overview for Cold Start

3 Data

The following section includes an overview of the source corpora as well as the training and evaluation data sets made available to TAC KBP performers in 2014. Tables listing all of these data sets are included as appendices at the end of this document.

3.1 Source Data

Source corpora for all of the TAC KBP 2014 efforts except Event Argument Extraction were distributed via three separate packages. The primary English sources used by most of the evaluations (regular Slot Filling, Sentiment Slot Filling, Entity Linking, Entity Discovery & Linking, and Cold Start) were repackaged and distributed as TAC 2014 KBP English Source Corpus (LDC2014E13). Although this package did not include any new source documents (all had been distributed in 2013 as part of the collected 2013 corpus - LDC2013E45), it did include output from the SERIF information extraction tool (Ramshaw, et al, 2011) on the full contents of the English sources, which included entity and relation extractions and syntactic parses, all of which were generously provided by BBN.

The source documents used to support the 2014 cross-lingual Entity Linking evaluations were distributed as TAC 2014 KBP Chinese Source Corpus (LDC2014E29) and TAC 2014 KBP Spanish Source Corpus (LDC2014E30). As with English, the Chinese corpus did not include any new documents, but did include SERIF output. Conversely, the Spanish data did include new data - 650,000 discussion forum threads taken from a collection developed for the DEFT program (LDC2014E14) - but did not include any SERIF output. This is because BBN felt that the system was not sufficiently mature in handling Spanish data to produce useful output given the current lack of document-level entity and relation training data.

Corpora	Genres	Documents
LDC2014E13: TAC 2014 KBP English Source Corpus	Newswire	1,000,257
	Web Text	999,999
	Discussion Forums	99,063
LDC2014E29: TAC 2014 KBP Chinese Source Corpus	Newswire	2,000,256
	Web Text	815,886
	Discussion Forums	199,321
LDC2014E30: TAC 2014 KBP Spanish Source Corpus	Newswire	910,734
	Discussion Forums	649,065

Table 7: Primary Source Corpora for TAC KBP 2014

The ‘official’ TAC KBP reference knowledge base (TAC KBP 2009 Evaluation Reference Knowledge Base - LDC2009E58), which had been used from 2009 through 2013 in the Entity Linking and Slot Filling evaluations, was only used by Entity Linking and its variation in 2014. The KB includes 818,741 nodes – articles drawn from an October 2008 dump of English Wikipedia – and each node corresponds to a unique entity corresponding to one of four types: person (PER), organization (ORG), geo-political entity (GPE), or unknown (UNK). All entries have semi-structured ‘infoboxes’, or tables of attributes pertaining to the subject entities. Some of the pages from the Wikipedia dump were not included in the KB because of ill-formatted infoboxes.

Two source corpora were selected for Event Argument Extraction, one for the pilot and one for the evaluation. While the pilot collection was simply a small subset of the 2014 English sources, the evaluation corpus was drawn from previously unreleased newswire and discussion forum threads.

3.2 Training and Evaluation Data

As 2013 marked LDC's sixth year of supporting the evaluations, performers participating in this year's TAC KBP were able to receive a wealth of materials for training and measuring the performance of their systems. Training data included forty-three datasets developed during the previous TAC KBP evaluation cycles - 14 for Entity Linking, 16 for regular Slot Filling, 5 for Temporal Slot Filling, 5 for Cold Start, and 3 for Sentiment Slot Filling. In addition to KBP-specific datasets, however, more data sets developed in support of other programs and evaluations were made available to KBP performers in 2014 than in previous years - 11 in total. Lastly, four new training corpora were developed in 2014 - pilot assessments for Event Argument Extraction, Chinese and Spanish Entity Linking data derived entirely from discussion forum threads, and the first set of Entity Discovery & Linking queries and KB links.

LDC produced 11 redistributable data sets for the 2014 evaluations, 2 each for Event Argument Extraction, Cold Start, Sentiment Slot Filling, and regular Slot Filling, and 1 each for Chinese and Spanish cross-lingual Entity Linking and Entity Discovery & Linking.

4 Annotation & Assessment Procedures and Methodologies

4.1 Cross-Lingual Entity Linking and Entity Discovery & Linking

Cross-Lingual Spanish and Chinese Entity Linking remained largely unchanged from 2013 to 2014 but a new variant of the English monolingual task – Entity Discovery & Linking - was performed for the first time in 2014. ED&L differed from Entity Linking in that source documents were exhaustively annotated according to modified Entity Linking guidelines, as opposed to 'cherry-picking' ambiguous entity mentions from the corpus as is done for regular EL.

The overall goals of query selection for cross-lingual Entity Linking did not change in 2014. As in previous years, annotators sought to collect the most ambiguous named entity mentions they could find in the corpus. Ambiguity was measured both by the number of distinct entities in the full query set referred to by the same name string (polysemy) as well as the number of distinct entities in the set that were referred to by multiple, unique named mentions (synonymy). For example, the string "Smith" would make a polysemous query because an annotator could probably find it in the corpus referring to different entities, while "Barack Obama" would make a synonymous query because the entity is also referred to in the corpus as "B. Hussein Obama" or "Bam Bam".

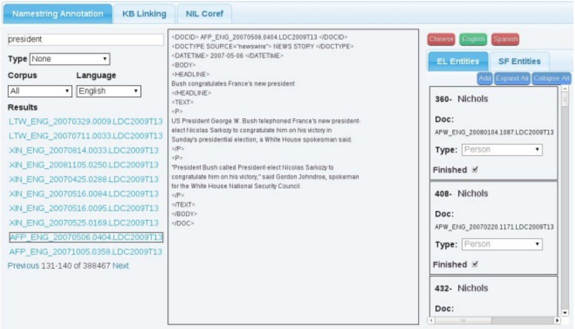
Cross-lingual Entity Linking queries were selected with the intention of representing as evenly as possible the three entity types

(PERs, ORGs, and GPEs) and the statuses of NIL (not linkable to any node in the KB) and non-NIL. In 2014, as opposed to previous years, each set of cross-lingual Entity Linking queries strove for a source document genre ratio of 1/2 newswire and 1/2 informal documents (discussion forum and web). For the cross-lingual versions of the task, although the majority of the queries were to be drawn from non-English documents, mentions in English documents of entities co-referential with other non-English queries were often selected.

In support of the new Entity Discovery & Linking task, annotators exhaustively annotated all named mentions of PERs, ORGs, and GPEs that occurred in documents containing at least 2-3 of the ambiguous entity mention types targeted for regular EL. This meant it was not possible to balance the ED&L data sets by entity type or NIL status, as the nature of exhaustive whole-document annotation does not allow for the necessary measure of control. ED&L queries were, however, balanced in terms of source document genre: roughly 1/3 of the queries were drawn from newswire documents, 1/3 from discussion forum threads, and 1/3 from web documents.

While all EL queries are either linked to the KB or marked NIL, a third category for entities – NIL Unknown – needed to be added to the task in order to support ED&L. Since annotators were required to capture all named mentions in a document, they could no longer avoid mentions that were impossible to confidently disambiguate. For instance, since post author usernames are pseudonyms, it is possible that a post author could in fact be an entity with a page in the knowledge base. However, as few personal details about post authors are normally revealed in threads, such determinations cannot be made, distinguishing them from truly NIL queries.

To select queries for both EL and ED&L, annotators searched the corpus, sometimes utilizing tagger output as a guide, and annotated any named entity mentions fitting the guidelines. Although annotators were not restricted to selecting queries from NER output, it was utilized by annotators while searching for polysemous strings to guide them through the corpus. However, in searching for synonymous entities, annotators' creativity, world knowledge, and research skills were the most effective tools.



president

Type None

Corpus All

Language English

Results

LTW_ENG_20070329.0009.LDC2009T13
LTW_ENG_20070711.0033.LDC2009T13
XIN_ENG_20070814.0033.LDC2009T13
XIN_ENG_20081105.0250.LDC2009T13
XIN_ENG_20070425.0288.LDC2009T13
XIN_ENG_20070516.0084.LDC2009T13
XIN_ENG_20070516.0095.LDC2009T13
XIN_ENG_20070525.0169.LDC2009T13
AFP_ENG_20070506.0404.LDC2009T13
AFP_ENG_20071005.0358.LDC2009T13
Previous 131-140 of 388467 Next

```
<DOCID> AFP_ENG_20070506.0404.LDC2009T13 </DOCID>
<DOCTYPE SOURCE="newswire"> NEWS STORY </DOCTYPE>
<DATETIME> 2007-05-06 </DATETIME>
<BODY>
<HEADLINE>
Bush congratulates France's new president
</HEADLINE>
<TEXT>
<P>
US President George W. Bush telephoned France's new president-
elect Nicolas Sarkozy to congratulate him on his victory in
Sunday's presidential election, a White House spokesman said.
</P>
<P>
"President Bush called President-elect Nicolas Sarkozy to
congratulate him on his victory," said Gordon Johndroe, spokeman
for the White House National Security Council.
</P>
</TEXT>
</BODY>
</DOC>
```

360- Nichols
Doc: APW_ENG_20080104.1087.LDC2009T13
Type: Person
Finished ☑

408- Nichols
Doc: APW_ENG_20070220.1171.LDC2009T13
Type: Person
Finished ☑

432- Nichols
Doc:

Figure 1: Namestring Annotation View of the EL and ED&L Query Development Tool

Although there are three distinct annotation phases to both EL and ED&L query development - namestring selection, knowledge base linking, and NIL co-reference, LDC's online interface allowed annotators to move back and forth between the three phases in order to more easily balance desired ratios of NIL and non-NIL queries (for EL) and to break up, and thereby simplify, NIL co-reference for both tasks.

4.2 Regular Slot Filling, Sentiment Slot Filling, and Cold Start

By design, there is a great deal of similarity and overlap between regular Slot Filling (SF), Sentiment Slot Filling (SSF), and Cold Start (CS) primarily due to the fact that all three tasks are scored by assessment of pooled responses that include a human-generated 'manual run'. While there are certainly differences between the tasks, which will be highlighted below, we will detail the processes for each collectively, both to avoid redundancy and to highlight subtle differences.

4.2.1 Changes for 2014

Across all three tasks – SF, SSF, and CS – justification, or minimum extents of provenance supporting the validity of a KBP relation, was altered in 2014. Justification was first added to Slot Filling in 2012 in an attempt to have systems and annotators highlight the sources of their assertions and, thereby, reduce the effort required for assessment by no longer requiring judges to review whole source documents for support. In 2012, justification was a single, minimal text extent proving the connection between the subject entity, via the selected slot, to the object entity, value, or string. In practice, however, the restriction to a single string often caused provenance to include lengthy portions of unrelated text. As a result, justification was altered in 2013 to allow for up to two discontiguous strings. Justification was altered further in 2014 to allow for up to four discontiguous strings that could be selected from as many separate documents. This facilitated a greater potential for inferred relations that would be difficult to justify with a single document. For instance, the following relation:

<Sheila Lukins -
per:countries_of_residence -
United States>

Might not be supported in a single document but could be justified by the following four text extents, each from a different document:

- *Sheila Lukins died Sunday at her Manhattan home*
- *Manhattan, the most densely populated of NYC's five boroughs*
- *New York, New York*
- *New York was the first US state to require vehicles to have license plates*

For the first time, ambiguous Slot Filling and Cold Start queries were developed in 2014. In previous evaluations, a candidate query was considered unambiguous and, thereby, appropriate for the task if its name string could be considered canonical (i.e. appropriate for use as the title of a Wikipedia page) and its entity referent could be easily identified by surrounding context. In 2014, ten ambiguous named entity mentions were selected for use as SF queries (the documents from which these queries were drawn were also then exhaustively annotated for the Entity Discovery & Linking task).

Lastly, 2014 was the first year in which the NIL and non-NIL status for Slot Filling queries was decided with respect to live Wikipedia instead of the official, static TAC KBP knowledge base. Using live Wikipedia in 2014 meant that NIL queries were more difficult for annotators to find, but also more challenging for systems to learn about.

4.2.2 Query Development

Much like EL and ED&L, query development for both of the Slot Filling task varieties and Cold Start was driven by guided searches through the corpus. Unlike the other evaluations, however, initial searches usually focused on key words related to the KBP slots for the given task, rather than an entity mention. For example, annotators might have searched for “arrested” or “charged” to develop queries that would generate fillers for the *per:charges* slot. Once an initial ‘seed’ annotation such as the above was found, query developers searched for other mentions of the connected entity or entities in the corpus to get a sense of how productive the query would be. Note, however, that while highly-productive queries were always desired, less productive queries were also selected if they offered opportunities to fill under-utilized slots or slot types.

Task-specific selection criteria were also considered during the query selection processes. The full set of Slot Filling queries was selected with the goal of representing approximately equally the three varieties of query types, namely, those that take named entities as fillers, those that take values as fillers (dates and numbers), and those that take strings as fillers. In the 2009-2012 Slot Filling evaluations, responses that were redundant with those already included in the official KB were marked as such and counted against scores, which led query developers to avoid non-NIL entities with fully fleshed out KB nodes. However, scoring was altered in 2013 (and again in 2014) such that redundant responses would neither negatively or positively impact scores. As a result, while query developers still largely avoided entities with fully-fleshed out KBs, greater flexibility was allowed in the more recent datasets.

Sentiment Slot Filling deviates from regular Slot Filling by specifying the slot to be filled as part of the queries, allowing for easier control of equal slot representation. Accordingly, SSF query developers could focus on selecting queries capable of generating edge-case or interesting fillers, and often both. For example, correctly extracting a response from the first of the following two statements is more challenging due to the inference needed to derive the sentiment from the stated desired action:

- *I think Michael Vick should have been executed for that”, said Carlson*
- *Carlson said he hated Michael Vick*

For the 2014 Sentiment Slot Filling evaluation, answers for a given query were drawn only from the same source document as the query, not the entire KBP corpus as had been done in 2013. Because of this need, query developers reviewed each document thoroughly to ensure that the majority of query documents contained a variety of answers for their relevant queries. However, some documents were also chosen such that no correct answers could be found for a certain query/document combination.

Cold Start query developers in 2014 searched through the TAC KBP English source corpus and looked for entities richly connected to others via KBP slot relations. For example, given the following text extent:

> *Jane Doe is the president of the School of Arts and Sciences at the University of Pennsylvania*

annotators could create the following query:

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

Note that, while the example above only lists a single filler for each of the two slots *(per:employee_of* and *org:parents*), there could potentially be multiple fillers at each of these “hop” levels, all of which had to be annotated and correctly connected to one another, a complication that sets Cold Start apart from the other Slot Filling tasks.

Validity decisions for Cold Start fillers were based on the same slot descriptions used for regular Slot Filling. However, in an attempt to increase connectivity between entities in the Cold Start corpus, inverses of all the existing Slot Filling slots were created for use in Cold Start. For example, for the existing slot *per:employee_or_member_of*, which captures organizations with which a person entity is affiliated as a member or employee, the inverse slot *org:employees_or_members* is used in order to also capture people who were affiliated with an organization entity.

4.2.3 Manual Runs

LDC developed manual runs, or the human-produced set of responses for each of the evaluation queries for regular and Sentiment Slot Filling as well as Cold Start. For each SF query, annotators were given up to two hours to search the corpus and locate all valid fillers. Since the set of responses for SSF could only be drawn from the query’s source document, less time was spent on producing the manual run for the task and more of the time allocated for the task was devoted to query development. Following the first passes of annotation for both tasks, quality control passes were conducted to flag any fillers that did not have adequate justification in the source document, or that might have been at variance with the current guidelines. These flagged fillers were then adjudicated by senior annotators and corrected or removed as appropriate.

As was done in all previous regular Slot Filling evaluations, slots from the Wikipedia infoboxes to which entities were linked during query development were mapped to one or more of the TAC KBP slots. Additionally, the existing fillers in the KB were edited and then loaded into the assessment tool to be available for coreference with responses generated during the evaluation (which, thereby, were marked as redundant). For example, if a given PER entity had “Philadelphia, PA” as its listed location of death in Wikipedia, that information would be separated into two filler strings (“Philadelphia” and “Pennsylvania”) and mapped to the KBP slots *per:city_of_death* and *per:stateorprovince_of_death*.

4.2.4 Assessment

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

From an assessor’s perspective, the Slot Filling and Cold Start assessment tasks were nearly identical except for some of the variations between the slots used. Fillers were marked as correct if they were found to be in-line with the slot descriptions and supported in the provided justification string(s) and/or its surrounding content. Fillers were assessed either as wrong if they did not meet both of the conditions for correctness or inexact if overly insufficient or extraneous text had been selected for an otherwise correct response. Justification was assessed as correct if it succinctly and completely supported the relation, wrong if it did not support the relation at all (or if the corresponding filler was marked wrong), inexact-short if part but not all of the information necessary to support the relation was provided, or inexact-long if it contained all information necessary to support the relation but also a great deal of extraneous text. Responses with justification comprised of more than 600 characters in total were automatically marked as ignored and given no assessment.

As with the development of the manual runs, after first passes of assessment were completed, quality control was performed on the data by annotators who reviewed the work of their peers and flagged potentially problematic assessments for additional review. As with the Slot Filling quality control procedure, this process improved assessment results while also indicating potential improvements in the guidelines and areas in which some annotators required more training.

4.2.5 Scores

The scores LDC receives on its manual runs help to identify when guidelines or other forms of annotator training or testing may be in need of improvement, in addition to indicating how well systems are faring against human in extracting information from text. Below are scores for LDC’s manual runs in the Slot Filling, Sentiment Slot Filling and Cold Start evaluations. Table 1 lists the results for both the 2013 and 2014 runs, for easy comparison (note, however, that no scores are available for

2013 Cold Start as the scorer required a knowledge base as input).

Track	Precision	Recall	F1
2013 Slot Filling	86%	57%	68%
2014 Slot Filling	88%	59%	70%
2013 Sentiment SF	70%	76%	73%
2014 Sentiment SF	86%	70%	77%
2014 Cold Start	91%	46%	62%

Table 8: LDC's Scores for Slot Filling, Sentiment Slot Filling, and Cold Start

LDC's regular Slot Filling scores in 2013 and 2014 were notably consistent, which we attribute largely to the task remaining relatively stable from 2013 to 2014 and, even more importantly, the task definition reaching a point at which query development could begin relatively early.

In a similar vein, it is worth pointing out the significant increase in the precision score for Sentiment Slot Filling. 2013 was the first year for Sentiment Slot Filling, and it is not surprising that, as such, LDC's precision in the task (70%) was significantly lower than our precision in other tasks that year. However, precision increased by 16 points (to 86%) in 2014, effectively rising to the level of our precision in Slot Filling in the 2013 and 2014 evaluations. We generally attribute this to the fact that, while certain aspects of the task design changed, the definitions of sentiment itself and the sentiment slots did not change, making the task virtually identical for LDC annotators.

The primary way in which Sentiment Slot Filling changed in 2014 was in the fact that, as discussed in section 4.2.2 above, fillers for a query were pulled from only a single document, whereas in 2013 annotators (and systems) searched for fillers in the entire corpus. Because of this single document limitation, we were surprised to see that LDC's recall dropped to 70% this year as our expectation was that having to review only a single document would mean higher recall in 2014. We plan to investigate further what might have led to this unexpectedly low recall and, although we expect that some assessment errors will be a factor, we are curious whether the power of suggestion may have played a role. That is, while LDC annotators may have been quite strict during the annotation phase, having answers presented to them as possibilities during assessment may have caused them to be more lenient. If this is the case, some number of system answers may have subsequently been marked correct even though annotators for the manual run had deemed them incorrect.

While LDC's precision in Cold Start in 2014 was the highest of the three tracks, recall was only 46%. While lower recall is not entirely unexpected, we were expecting a precision score closer to that of regular Slot Filling and so also plan to investigate the roles that the power of suggestions and annotation or assessment errors played in our results.

4.3 Event Argument Extraction

In Event Argument Extraction, a new task in 2014, annotators and systems extracted mentions of entities from unstructured text and indicated the role they played in an event as supported by text. Critically, the extracted information had to be suitable as input to a knowledge base and so annotators and systems produced tuples indicating the event type, the role played by the entity in the event, and the most canonical mention of the entity itself from the source document. Event Argument Extraction 2014 was made up of three separate processes – source document selection, manual run development, and assessment.

4.3.1 Document Selection

Documents served as queries in EAE and so the first step for annotators in developing data for the task was to perform targeted searches over two sets of previously unreleased documents (one set of newswire documents and one set of discussion forum threads). Documents were selected based on the criteria that they contained at least one mention of the specified event types along with valid arguments for the event. Documents with a variety of event types were primarily sought after, though documents providing mentions of generally less common event types were also selected for inclusion.

Upon finding a promising document, selectors reviewed the text closely and tallied the number of unique event mentions of each event type that was included. Such tallies helped ensure that all of the targeted event types were at least reasonably well-represented in the corpus of documents selected for the Event Argument Extraction evaluation. While performing document reviews, annotators also searched for certain undesirable qualities that would prevent a document from being included in the corpus. Most notably, discussion forum documents with more than a small amount of newswire quotation were avoided with the aim of selecting discussion forum data actually comprised of informal content.

4.3.2 Manual Run

For each document in the EAE evaluation, an annotator had only a maximum of thirty minutes to read through the text and annotate all valid, unique event arguments within that document. As with Slot Filling, Sentiment Slot Filling and Cold Start, following the initial round of annotation, a quality control pass was conducted to flag any event arguments that did not have adequate justification in the source document, or that might be at variance with the current guidelines. These flagged annotations were then adjudicated by senior annotators.

4.3.3 Assessment

For the assessment of EAE responses produced during the evaluation, LDC used an online tool developed and graciously provided by BBN. The decision was made to switch from the LDC-developed assessment tool used for the pilot assessment because, while the pilot pointed to some updated features that could enhance the LDC tool, other competing TAC KBP technical needs complicated completion of the upgrades while the BBN tool was readily available.

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

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

As assessment was completed, quality control was performed on the data using a procedure similar to that described above for other tasks. Senior annotators reviewed the work of assessors and made corrections to assessment kits and, for each correction that was made, the reviewer followed up with the original assessor to clarify the correction. Given the newness and complexity of the task, weekly meetings were also held in addition to peer reviews in order to allow assessors to discuss difficult examples they had encountered and come to a consensus regarding how to handle them. These meetings helped ensure that less well-defined areas of the guidelines were being handled consistently across assessors.

4.3.4 Scores

Table 2 provides the results of LDC's manual run produced for the Event Argument Extraction track. These scores resulted from a preliminary (partial) release containing LDC's assessments of the responses for approximately one-fourth of the documents annotated in Event Argument Extraction. As such, it is possible that these scores will change to some degree as additional assessments are completed.

Track	**Precision**	**Recall**	**F1**
Event Argument Extraction	77%	29%	42%

Table 9: LDC's Event Argument Extraction Scores

Given that Event Argument Extraction was a new task in 2014, it is not surprising that LDC's precision for the track is lower than that for Slot Filling, Sentiment Slot Filling and Cold Start. However, there are a few other important differences between EAE and the other tasks that we believe impacted the score. Primarily, whereas each response for the other three tracks received only two assessments, six judgments were made for each EAE response, five of which could have been points of disagreement between the annotator who produced the manual run and the assessor. Considering the greater complexity of the task, and the fact that assessors encountered phenomena in the data not predicted in task guidelines (due, again, to the newness of the task), 77% precision is not beneath expectations. Compared with LDC's precision in the first year of Sentiment Slot Filling, a notably less complex task, precision for EAE was seven points higher this year.

Perhaps what might be more surprising on the surface is LDC's recall, only 29%, given that EAE responses were drawn from a single document rather than an entire corpus. Be that as it may, we believe that two important differences between EAE and other KBP tasks with assessments led to the low recall.

First and foremost was the restriction that, during the annotation phase, annotators could spend no more than 30 minutes on each document. While this was generally enough time to complete a surface-level first pass of a document, it did not provide annotators with the time needed to think about the event arguments in a document beyond that linear, surface interpretation. As a result annotators did not capture many of the potential inferred event arguments found by systems.

Second, systems were far more exhaustive than annotators in capturing nominal phrases. For instance, given the phrase "some of the men arrested", annotators were more inclined to annotate only a single *Justice.Arrest-Jail Person* argument: "the men arrested". Systems, on the other hand, were more likely to capture two arguments: "the men arrested" and "some of the men arrested" and, since these two mentions are not co-referential, they led to two, unique correct answers. Additionally, the instruction that annotators for the manual run should only return unique responses may have had an effect on the score as well; since "the men arrested" is a superset of "some of the men arrested", annotating only the former might have felt intuitively more complete to a rushed annotator. Lastly, the varying ways in which systems returned nominal phrases often caused assessors to place very similar arguments in separate equivalence classes, increasing the number of event arguments that were missed by LDC annotators.

5 Conclusion

This paper discussed the linguistic resources produced in support of the TAC KBP 2014 evaluations, from the planning processes and data creation efforts to descriptions of the datasets and analysis of how results compared to previous efforts. LDC support of TAC KBP in 2014 included contributions to task descriptions, data curation and distribution, source corpus expansion, and creating or revising existing data development procedures to accommodate new or modified evaluations. Future work will include repackaging and updating documentation to make the data created this year more readily useable in the future by system developers who may be unfamiliar with the KBP evaluations. The resources described in this paper are slated for publication in the LDC Catalog, in order to make the corpora available to the wider research community. Other resources such as KBP system descriptions and site papers will be published on the NIST TAC website.

6 References

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

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

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

Paul McNamee, Hoa T. Dang, Heather Simpson, Patrick Schone, and Stephanie M. Strassel. 2010. An Evaluation of Technologies for Knowledge Base Population. In *Proceedings of the Seventh International Language Resources and Evaluation Conference (LREC)*, Valleta, Malta.

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

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

Appendix A: Data Available to KBP Performers in 2014

Table 1: 2009 – 2013 Entity Linking Data Sets

Corpus Title	LDC Catalog	Language	Size
TAC 2009 KBP Gold Standard Entity Linking Entity Type List	LDC2009E86	English	567 GPE
			627 PER
			2710 ORG
TAC 2010 KBP Evaluation Entity Linking Gold Standard	LDC2010E82	English	749 GPE
			741 PER
			750 ORG
TAC 2010 KBP Training Entity Linking	LDC2010E31	English	500 GPE
			500 PER
			500 ORG
TAC 2011 KBP Cross-lingual Training Entity Linking	LDC2011E55	Chinese English	685 GPE
			817 PER
			660 ORG
TAC 2011 KBP English Evaluation Entity Linking Annotation v1.1	LDC2011R36	English	750 GPE
			750 PER
			750 ORG
TAC 2011 KBP Cross-lingual Evaluation Entity Linking Annotation V1.1	LDC2011R38	Chinese English	642 GPE
			824 PER
			710 ORG
TAC 2012 KBP Chinese Entity Linking Evaluation Annotations	LDC2012E103	Chinese English	605 GPE
			699 PER
			718 ORG
TAC 2012 KBP Chinese Entity Linking Web Training Queries and Annotations	LDC2012E66	Chinese English	52 GPE
			52 PER
			54 ORG
TAC 2012 KBP English Entity Linking Evaluation Annotations	LDC2012E102	English	604 GPE
			919 PER
			706 ORG
TAC 2012 KBP Spanish Entity Linking Evaluation Annotations	LDC2012E101	Spanish English	858 GPE
			669 PER
			539 ORG
TAC 2012 KBP Spanish Entity Linking Training Queries and Annotations	LDC2012E67	Spanish English	566 GPE
			683 PER
			601 ORG
TAC 2013 KBP English Entity Linking Evaluation Queries and Knowledge Base Links	LDC2013E90	English	803 GPE
			686 PER

			701 ORG
TAC 2013 KBP Chinese Entity Linking Evaluation Queries and Knowledge Base Links	LDC2013E96	Chinese English	714 GPE
			706 PER
			735 ORG
TAC 2013 KBP Spanish Entity Linking Evaluation Queries and Knowledge Base Links	LDC2013E97	Spanish English	660 GPE
			695 PER
			762 ORG

Table 2: 2009 – 2013 Regular Slot Filling Data Sets

Corpus Title	LDC Catalog	Language	Size
TAC KBP 2009 Evaluation Slot Filling List	LDC2009E65	English	53 Queries
TAC KBP 2009 Assessment Results	LDC2009E90	English	10,416 Assessments
TAC 2010 KBP Training Slot Filling Annotation	LDC2010E18	English	50 Queries
TAC 2010 KBP Evaluation Slot Filling Annotation	LDC2010R11	English	100 Queries
TAC 2010 KBP Assessment Results	LDC2010E61	English	25,511 Assessments
TAC 2010 KBP Training Surprise Slot Filling Annotation	LDC2010E52	English	32 Queries
TAC 2010 KBP Evaluation Surprise Slot Filling Annotation	LDC2010E52	English	40 Queries
TAC 2011 KBP English Training Regular Slot Filling Annotation	LDC2011E48	English	48 Queries
TAC 2011 KBP English Evaluation Regular Slot Filling Annotation V1.2	LDC2011E89	English	100
TAC 2011 KBP English Regular Slot Filling Assessment Results V1.2	LDC2011E88	English	28,041 Assessments
TAC 2012 KBP English Regular Slot Filling Evaluation Annotations V1.1	LDC2012E91	English	80 Queries
TAC 2012 KBP English Regular Slot Filling Assessment Results V1.2	LDC2012E115	English	22,885 Assessments
TAC 2012 KBP Spanish Slot Filling Training Queries and Annotations V1.2	LDC2012E68	Spanish English	50 Queries
TAC 2013 English Regular Slot Filling per:title Training Data	LDC2013E60	English	1949 Assessments

TAC 2013 English Regular Slot Filling Evaluation Queries and Annotations	LDC2013E77	English	100 Queries
TAC 2013 English Regular Slot Filling Evaluation Assessment Results V1.1	LDC2013E91	English	27,655 Assessments

Table 3: 2011 – 2013 Temporal Slot Filling Data Sets

Corpus Title	LDC Catalog	Language	Size
TAC 2011 KBP English Training Temporal Slot Filling Annotation	LDC2011E49	English	50 Queries
TAC 2011 KBP English Evaluation Temporal Slot Filling Annotation	LDC2012E38	English	100 Queries
TAC 2013 KBP English Temporal Slot Filling Training Queries and Annotations	LDC2013E82	English	7 Queries
TAC 2013 KBP English Temporal Slot Filling Evaluation Queries and Annotations	LDC2013E86	English	273 Queries
TAC 2013 KBP English Temporal Slot Filling Evaluation Assessment Results	LDC2013E99	English	4,376 Assessments

Table 4: 2012 – 2013 Cold Start Data Sets

Corpus Title	LDC Catalog	Language	Size
TAC 2012 KBP Cold Start Queries V1.1	LDC2012E105	English	385 Queries
TAC 2012 KBP Cold Start Assessment Results	LDC2012E116	English	5015 Assessments
TAC 2012 KBP Cold Start Automated Queries Assessment Results	LDC2013E39	English	779 Assessments
TAC 2013 KBP English Cold Start Evaluation Queries and Annotations V1.1	LDC2013E87	English	326 Queries
TAC 2013 KBP English Cold Start Evaluation Assessment Results	LDC2013E101	English	6,755 Assessments

Table 5: 2013 Sentiment Slot Filling Data Sets

Corpus Title	LDC Catalog	Language	Size
TAC 2013 KBP English Sentiment Slot Filling Training Queries and Annotations	LDC2013E78	English	160 Queries
TAC 2013 KBP English Sentiment Slot Filling Evaluation Queries and Annotations	LDC2013E89	English	160 Queries

TAC 2013 KBP English Sentiment Slot Filling Evaluation Assessment Results	LDC2013E100	English	5,160 Assessments

Table 6: 2014 Training Data

Corpus Title	LDC Catalog	Language	Size
TAC 2014 KBP English Entity Discovery and Linking Training Data V1.3	LDC2014E54	English	5786 queries
TAC 2014 KBP Spanish Entity Linking Discussion Forum Training Data V1.1	LDC2014E46	Spanish English	541 queries
TAC 2014 KBP Chinese Entity Linking Discussion Forum Training Data	LDC2014E47	Chinese English	514 queries
TAC 2014 KBP Event Argument Extraction Pilot Assessment Results V1.1	LDC2014E40	English	11,625 assessments

Table 7: 2014 Evaluation Data

Corpus Title	LDC Catalog	Language	Size
TAC 2014 KBP English Regular Slot Filling Evaluation Queries and Annotations V1.1	LDC2014E66	English	100 queries
TAC 2014 KBP English Regular Slot Filling Evaluation Assessment Results V2.0	LDC2014E75	English	21,956 assessments
TAC 2014 KBP Chinese Entity Linking Evaluation Queries and Knowledge Base Links V2.0	LDC2014E83	Chinese English	2739 queries
TAC 2014 KBP Spanish Entity Linking Evaluation Queries and Knowledge Base Links	LDC2014E84	Spanish English	2057 queries
TAC 2014 KBP English Entity Discovery and Linking Evaluation Queries and Knowledge Base Links V1.1	LDC2014E81	English	5234 queries
TAC 2014 KBP English Sentiment Slot Filling Evaluation Queries and Annotations V1.1	LDC2014E72	English	400 queries
TAC 2014 KBP English Sentiment Slot Filling Evaluation Assessment Results	LDC2014E85	English	6,383 assessments
TAC 2014 KBP English Cold Start Evaluation Queries and Annotations V1.1	LDC2014E73	English	247 queries

TAC 2014 KBP English Cold Start Evaluation Assessment Results V2.0	LDC2014E82	English	7258 assessments
TAC 2014 KBP English Event Argument Extraction Evaluation Annotations V1.1	LDC2014E74	English	5947 annotations
TAC 2014 KBP English Event Argument Extraction Evaluation Assessment Results V1.1	LDC2014E88	English	18,878 assessments