Linguistic Resources for 2012 Knowledge Base Population Evaluations

Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M. Strassel, & Jonathan Wright

Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104
U.S.A

Email: {joellis, xuansong, kiragrif, strassel, jdwright} @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). The KBP evaluation includes three tasks that target information extraction and question answering technologies: Entity Linking, Slot Filling, and Cold Start. The Cold Start task was introduced in 2012 in an effort to combine and enhance technologies developed for Slot Filling and Entity Linking. Linguistic Data Consortium (LDC) has supported the TAC KBP evaluation since 2009, each year producing new linguistic resources including data, annotations, system assessments, tools and specifications. This paper describes the resource creation efforts in support of TAC KBP 2012, with an emphasis on procedures and methodologies for query selection, annotation, and assessment.

1. Introduction

The Text Analysis Conference (TAC) is a series of evaluation workshops initiated by the National Institute of Standards and Technology (NIST) that aim to advance natural language processing technologies and applications. Knowledge Base Population (KBP), one of the on-going TAC tracks, started in 2009 with a focus on information extraction and question answering technologies. Evolved from the TREC Question Answering (Dang et al. 2006) and Automated Content Extraction (ACE) (Doddington et al. 2004) evaluation programs (McNamee et al. 2010), TAC KBP evaluates computation systems on three main tasks: Entity Linking, Slot Filling, and Cold Start.

The Entity Linking task requires systems to either accurately link named mentions of person (PER), organization (ORG), and geopolitical (GPE) entities in text to entries in an external knowledge base, or correctly report if there are no matching entries. Entity Linking evaluations started in 2009 with an English only version (Simpson et al., 2010) and added a Chinese cross-lingual version of the task in 2011. In 2012, cross-lingual Entity Linking evaluations were expanded with a Spanish version of the task. The Slot Filling task requires systems to automatically populate Wikipedia-style infoboxes for a set of specific named person (PER), and organization (ORG) entities with information retrieved from a collection of natural language English source documents. In 2012, the addition of Spanish Slot Filling moved the task into cross-lingual terrain. Cold Start, a new KBP task created in 2012, requires systems to construct a new knowledge base from the information contained in an unstructured text collection, effectively coordinating the separate technologies developed for Entity Linking and Slot Filling.

Linguistic Data Consortium (LDC) at the University of Pennsylvania has supported KBP evaluations since 2009 by creating and distributing linguistic resources including data, annotations, system assessment, tools and specifications. This paper describes the resource creation effort for 2012 TAC KBP. Section 2 describes the source data and knowledge base used for all KBP tasks; section 3 discusses the training and evaluation data provided by LDC for the 2012 KBP tasks; section 4 discusses procedures and methodologies for query selection, annotation, and assessment; and section 5 concludes the paper.

2. Source Data & Reference Knowledge Base

In 2011, the combined source corpus for all KBP tasks consisted of approximately 2.7 million English and Chinese documents from newswire, web, and other sources. Before the start of data creation efforts for 2012, the size of combined corpus for the Entity Linking and Slot Filling evaluations nearly tripled after the addition of 4.8 million new web and newswire documents. These additions were made in order to broaden the overall epoch covered by the corpus; add Spanish documents to the pool; create greater overlap between the English, Chinese, and Spanish sets; expand on the web document component of the collection; and ease the creation of unique queries for all tasks. Table 1 provides a breakdown of the documents currently included in the collection (see section 4 for a discussion of source corpus used for the Cold Start task in 2012).

All English documents included in the KBP source corpus prior to 2012 can be found in TAC 2010 KBP Source Data (LDC2010E12). The documents in this collection continued to be used for query development and annotations for the 2012 Entity Linking and Slot Filling evaluations, including the cross-lingual versions of these tasks. All of the new English, Chinese, and Spanish newswire documents added to the corpus in 2012 were drawn, respectively, from English Gigaword Fifth Edition (LDC2011T07), Chinese Gigaword Fifth Edition (LDC2011T13), and Spanish Gigaword Third Edition (LDC2011T12). The lists of documents selected from the Gigaword collections are included in TAC KBP 2012 Newswire Source Corpus Additions V1.1 (LDC2012E22). All new English and Chinese web documents that were added to the KBP source corpus in 2012, which were drawn from various collections previously compiled for the GALE project, can be found in TAC 2012 KBP Source Corpus Additions Web Documents (LDC2012E23).

Language	Genre	Documents
English	Broadcast Conversation	17
	Broadcast News	665
	Conversation Telephone Speech	1
	Newswire (2007 – 2010)	2,286,866
	Web Text (2008 – 2009)	1,490,595
Chinese	Newswire (1991 – 2010)	2,000,256
	Web Text (1997 – 2009)	815,886
Spanish	Newswire (2007 – 2010)	1,000,020

Table 1: 2012 Document Source Collection for Entity Linking and Slot Filling Tasks

The reference knowledge base (KB) (LDC2009E58) used in both the Entity Linking and Slot Filling tasks includes 818,741 nodes – articles drawn from an October 2008 dump of English Wikipedia. Each node corresponds to a unique entity corresponding to one of four types: person (PER), organization (ORG), geopolitical-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.

3. Training and Evaluation Data

As 2012 marked LDC's fourth year of supporting KBP evaluations, system developers participating in this year's Entity Linking and Slot Filling evaluations were able to receive a wealth of materials for training their systems before the evaluations began. For Entity Linking, six corpora developed in previous years as either training or evaluation materials for English or the cross-lingual Chinese versions of the task were made available to registered participants. In addition, LDC developed two new Entity Linking training corpora in 2012 to assist in participants' preparations for handling Chinese web documents and the new Spanish newswire documents. Including the three new releases developed for the English, Chinese and Spanish evaluations, five new Entity Linking corpora were developed in 2012.

For Slot Filling, five corpora developed in previous years for regular English Slot Filling as well as the Temporal Slot Filling task from 2011 were made available to participants as training materials (not including assessment corpora). In addition, two new Slot Filling releases were made in 2012 for Spanish training materials and for the English evaluation. Queries and annotations for a Spanish evaluation were also produced but never released. For Cold Start, one new corpus was produced which will be available for participant training in the future.

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

Table 2: Entity Linking Annotation Data

Corpus Title	Type	LDC Catalog	Language	Size (Queries)
TAC 2010 KBP Training Slot Filling Annotation	Training	LDC2010E18	English	25 PER
				25 ORG
TAC 2010 KBP Evaluation Slot Filling Annotation	Evaluation	LDC2010R11	English	50 PER
				50 ORG
TAC 2011 KBP English Training Temporal Slot Filling Annotation	Training	LDC2011E49	English	40 PER
				10 ORG
TAC 2011 KBP English Evaluation Regular Slot Filling Annotation V1.2	Evaluation	LDC2011R34	English	50 PER
				50 ORG
TAC 2011 KBP English Evaluation Temporal Slot Filling Annotation	Evaluation	LDC2011R40	English	80 PER
				20 ORG
TAC 2012 KBP Spanish Slot Filling Training Queries and Annotations V1.2	Training	LDC2012E68	Spanish English	25 PER
				25 ORG
TAC 2012 KBP English Regular Slot Filling Evaluation Annotations V1.1	Evaluation	LDC2012E91	English	40 PER
				40 ORG
TAC 2012 KBP Cold Start Queries V1.1	Evaluation	LDC2012E105	English	385 Queries

Table 3: 2012 Slot Filling and Cold Start Training and Evaluation Data

Dataset	Language	KB link	GPE	ORG	PER
TAC 2012 KBP Chinese Entity Linking Evaluation Annotations	Chinese	Non-NIL: NW/WB	164/131	167/112	148/110
		NIL : NW/WB	99/88	89/86	167/68
	English	Non-NIL: NW/WB	101/26	107/52	83/39
		NIL: NW/WB	90/6	79/26	68/16
TAC 2012 KBP Chinese Entity Linking Web Training Queries and Annotations	Chinese	Non-NIL: WB	24	27	24
		NIL: WB	18	18	19
	English	Non-NIL: WB	7	5	9
		NIL: WB	3	4	0
TAC 2012 KBP Spanish Entity Linking Evaluation Annotations	Spanish	Non-NIL: NW	559	150	159
		NIL: NW	248	366	509
	English	Non-NIL: NW/WB	40/0	11/1	0/0
		NIL: NW/WB	8/3	11/0	0/1
TAC 2012 KBP Spanish Entity Linking Training Queries and Annotations	Spanish	Non-NIL: NW	417	245	255
		NIL: NW	103	218	230
	English	Non-NIL: NW	29	37	62
		NIL: NW	17	101	136
TAC 2012 KBP English Entity Linking Evaluation Annotations	English	Non-NIL: NW/WB	341/188	143/133	270/105
		NIL: NW/WB	41/34	245/185	433/111

Table 4: Entity Linking Query Proportion Distribution

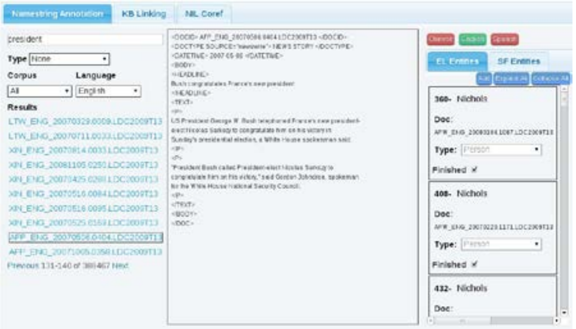
4. Annotation & Assessment Procedures and Methodologies

4.1 Entity Linking

The overall goals of query selection for Entity Linking did not change in 2012. As in previous years, annotators sought to collect the most confusable named entity mentions they could find in the corpus for use as training and evaluation queries. A query's confusability is 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). For example, the namestring "Smith" would be highly confusable because one could likely find numerous instances of it being used in the corpus to refer to different entities. Additionally, entities with numerous nicknames and shortened or misspelled versions of their names in the corpus were targeted to increase synonymy in the query set.

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 linked to the KB) and non-NIL. As was done in 2011, each set of Entity Linking queries strove for a source document genre ratio of 2/3 newswire to 1/3 web or informal documents. Lastly, 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 selected whenever possible (see Table 4).

Although the goals remained the same, the approach used to select Entity Linking queries was significantly altered in 2012. This change was largely made possible by a new Entity Selection tool developed by programmers at LDC. The primary advantage of the new tool was that it allowed annotators to search the corpus and select any text extent from source documents for use as queries. This was a major improvement over the method used from 2009 to 2011, in which query namestrings were restricted to those previously identified by an automated named entity tagger. Annotators could still utilize tagger output as a guide through the corpus but, once a confusable namestring was found, they could search for and annotate any other strings in the collection to maximize polysemy and synonymy and to balance distribution of source document genres and languages.



Namestring Annotation
KB Linking
NIL Coref
president
Type
None
Corpus
Language
All
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_20070525.0159.LDC2009T13
AFP_ENG_20070508.0404.LDC2009T13
AFP_ENG_20071005.0358.LDC2009T13
Previous 131-140 of 388467 Next
EL Entities
SF Entities
360- Nichols
Doc:
Type:
Person
Finished ✓
408- Nichols
Doc:
Type:
Person
Finished ✓
432- Nichols
Doc:

```
<DOCID> AFP_ENG_20070508.0404.LDC2009T13 </DOCID>
<DOCTYPE SOURCE="newswire"> NEWS STORY </DOCTYPE>
<DATETIME> 2007-05-08 </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, spokesman for the White House National Security Council.
</P>
</TEXT>
</BODY>
</DOC>
```

Figure 1: Namestring Annotation View of the Entity Selection Tool

Another major advantage of the new user interface was that it allowed for the three main phases of the Entity Linking task (namestring selection, KB linking, and NIL coreference) to be performed concurrently by annotators. Previously, programmer intervention was necessary to move data between the three phases, leading to a need for over selection of queries in order to end up with desired ratios. With the new tool, however, annotators could easily search the KB during namestring selection and determine whether a potential query would be NIL or non-NIL. The new interface also made NIL coreference easier as the task could be performed on reasonably sized batches rather than all at once at the end of the pipeline.

4.2 Slot Filling – Entity Selection

As was the case with Entity Linking, the goal of Entity Selection for the Slot Filling task remained unchanged in 2012 but a new GUI greatly eased and enhanced the process. Annotators performed guided searches through the corpus and selected mentions of entities based on three criteria: non-confusability, productivity, and uniqueness. A candidate query was considered non-confusable if its namestring could be considered full (i.e. appropriate for use as the title of a Wikipedia page) and its referent could be easily identified by surrounding context. Productivity for candidate queries was determined by whether the source corpus contained at least 2 - 3 slot fillers for the entity. Lastly, a potential Slot Filling query was said to be unique if the source corpus contained information on the entity that pertained to a KBP slot that had been under-utilized in previous evaluations. Targeted unique slots included:

per:cause_of_death
per:charges
org:political_and_religious_affiliations
org: number_of_employees_or_members
org:dissolved
org:website

The new entity selection tool for Slot Filling greatly eased annotators' efforts to meet these goals. The ability to perform searches across the corpus and capture any strings possible allowed them to use keywords related to the under-utilized KBP slots (e.g. "arrested" for *per:charges* or "died" for *per:cause_of_death*) and select entities connected to those phrases rather than having to rely on tagger output for finding such entities. Additionally, the new tool enabled annotators to refer to the KB during the namestring selection process to more easily balance NIL and non-NIL selections and to ensure that no entities with full KBs would be selected.

4.3 Slot Filling – Annotation

Preliminary steps to Slot Filling annotation included guidelines revisions, slot mapping for the selected entities that were linked to the KB, and annotator training. Based on annotator questions that arose during the 2011 evaluation, the descriptions of 15 slots were edited for clarity and to ensure greater continuity between training and assessment data. The Slot Filling guidelines were also altered in order to adapt to changes in the task requirements for 2012, which are described below.

As was done before previous Slot Filling evaluations, information from the Wikipedia infoboxes for entities linked to the KB during entity selection was mapped to one or more of the TAC KBP slots. For example, if a given PER entity had "Philadelphia, PA" as its listed "Death Location" 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:state_of_death.* Mappings were performed automatically and manually before results were reviewed and edited for consistency.

Potential Slot Filling annotators were provided with copies of the updated guidelines and a hands-on training session before being tested on their understanding of the slots and, thereby, their ability to successfully complete the task. This test consisted of 65 examples of varying degrees of difficulty, collected during the review of 2010 and 2011 Slot Filling data. Only annotators who successfully completed testing were able to participate in the Slot Filling annotation task.

Annotation was performed using LDC's Slot Filling GUI, which includes corpus search, annotation, and coreference components. For each query, annotators were given two hours in which to search the corpus and locate all valid fillers for the set of slots of their assigned entity. New to the annotation process in 2012, annotators were required to identify justification text extents for all selected fillers. Valid justification strings were said to clearly identify all three elements of a relation (i.e. the subject entity, the predicate slot, and the object filler), and the relation between them, with minimal extraneous text. Another change to the Slot Filling annotation process in 2012 was that annotators were instructed to capture and coreference duplicate fillers in order to provide more training data for systems.

Following the initial round of annotation, a quality control pass was conducted to flag any fillers that did not have adequate justification in the source document, or that might be at variance with the current guidelines. These flagged fillers were then adjudicated by senior annotators. This QC process was useful because in addition to providing a level of quality control it also provided information on areas of the guidelines in need of further clarification.

4.4 Slot Filling – Assessment

Preliminary steps for Slot Filling Assessment also included annotator training and guidelines revisions based on past lessons and the need to account for changes in the task. After an initial training session and guidelines review, candidate Slot Filling assessors were required to complete an assessment screening kit, which contained 12 filled slots for an actual entity. 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 the validity of slot-filling answers from both humans and systems and to create equivalence classes from fillers assessed as correct.

After assessment was completed, quality control was performed on the data using a procedure similar to that described above for slot filling annotation, in which annotators 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 deficiencies in the guidelines and areas in which some annotators required more training.

4.5 Cold Start – Corpus Selection

The first steps taken at LDC in preparation for the Cold Start evaluation involved discussions with coordinators to assist in task specifications and scouting for the evaluation corpus. Annotators searched online and reviewed dozens of different websites to determine suitability for the task. Ideal sites were made up of at least 10K documents and included numerous person (PER), organization (ORG), and geo-political entities (GPE) related to one another by the TAC KBP slots. After the site was selected, LDC's technical team worked with external KBP coordinators to harvest and process the documents.

4.6 Cold Start – Query Development

Following document processing, a named entity tagger was run on the Cold Start document collection. Annotators then performed a time-limited review of the tagger output, removing obviously bad elements in the list before it was given to task participants as 'entry points' into the collection.

Using an interface similar to the Entity Selection tool for the Entity Linking and Slot Filling tasks, annotators searched through the Cold Start corpus and looked for entities richly connected to others via KBP slot relations. For example, given the two following text extents:

> "Jane Doe is the president of the
> School of Arts and Sciences at the
> University of Pennsylvania"
>
> "The University of Pennsylvania,
> located in Philadelphia"

Annotators would create the following query and attempt to fill out each level of the query with all valid fillers for the entity/slot combination:

```
"Jane Doe"
    per:employee_of
    "School of Arts and Sciences"
            org:parents
            "University of Pennsylvania"
                    org:city_of_headquarters
                    "Philadelphia"
```

Validity decisions were based on the same slot descriptions used for the Slot Filling tasks. However, in an attempt to increase connectivity between entities in the Cold Start corpus, a few inverse versions of existing slots were created (e.g. for the existing slot *per:member_of*, which captures organizations with which the entity person is affiliated as a member, the inverse slot *org:members* was created to indicate people who were affiliated with the entity organization as members) (Cold Start Knowledge Base Population at TAC 2012 Task Description, 2012).

4.7 Cold Start – Assessment

The last stage of LDC's annotation efforts in support of the Cold Start task was to assess a subset of the contents of system generated KBs using the previously developed queries. From an annotator's perspective, Cold Start assessment was nearly identical to that of Slot Filling, except that only a single slot and set of fillers were assessed for each entity and fillers for the new inverse slots also had to be assessed.

5. Conclusion

This paper discussed procedures and methodologies for annotation and assessment for KBP 2012, particularly elaborating on procedures and methodologies for query selection, annotation, and assessment. LDC support of KBP in 2012 included source corpus expansion; significant revisions to the entity selection processes for both the Entity Linking and Slot Filling tasks in order to support coordinator requests for more challenging and diverse queries; revision of the annotation process, infrastructure, and data collected for Slot Filling; expansion of cross-lingual data with the addition of Spanish Entity Linking and Slot Filling; as well as the addition of a whole new task – Cold Start – which brought the total number of tasks supported to six in 2012, two more than in 2011. Future work will include further refinement of the changes made to tasks this year. 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.

References

Hoa T. Dang, Jimmy Lin, and Diane Kelly. 2006. Overview of the TREC 2006 Question Answering Track. In Proceedings of TREC 2006.

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.

Ralph Grishman, David Westbrook, and Adam Meyers. 2005. NYU's English ACE 2005 System Description. In Proceedings of the ACE 2005 Evaluation/PI Workshop.

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.

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 LREC.

Cold Start Knowledge Base Population at TAC 2012 Task Description. 2012. http://www.nist.gov/tac/2012/KBP/task_guidelines/Cold%20Start%202012%20Task%20Description%201.3.pdf