Linguistic Resources for 2011 Knowledge Base Population Evaluation

Xuansong Li, Joe Ellis, Kira Griffit, Stephanie M. Strassel
Robert Parker, Jonathan Wright

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

Email:{xuansong,joellis,kiragrif,strassel,parkerrl,jdwright}@ldc.upenn.edu

Abstract

The Knowledge Base Population (KBP) is an evaluation track of the Text Analysis Conference (TAC) workshop series organized by the National Institute of Standards and Technology (NIST). The KBP evaluation includes two tasks that target information extraction and question answering technologies: Entity Linking and Slot Filling. Cross-lingual Entity Linking and Temporal Slot Filling were introduced in 2011 to evaluate systems' abilities to recognize multilingual and temporal information. Linguistic Data Consortium (LDC) supports the TAC KBP evaluation by producing linguistic resources including data, annotations, system assessment, tools and specifications. This paper describes the resource creation efforts in support of KBP 2011, with an emphasis on annotation and assessment procedures and methodologies.

1 Introduction

The Text Analysis Conference (TAC) is a series of evaluation workshops initiated by the National Institute of Standards and Technology (NIST), aiming to advance natural language processing technologies and applications. The 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, the KBP track has evaluated computation systems on two main tasks: Entity Linking and Slot Filling (McNamee et al. 2010). The Entity Linking task requires systems to either accurately link mentions of person (PER), organization (ORG), and geopolitical (GPE) names in text to entries in an external knowledge base, or correctly report if there are no matched entries. The evaluation started in 2009 with English only (Simpson et al., 2010) and added Chinese for a cross-lingual task in 2011. The Slot Filling task requires systems to automatically populate Wikipedia-style infoboxes for a set of specific named entities with information retrieved from a collection of unstructured English source documents. In 2011, KBP created a new task by adding a temporal component to Slot Filling, requiring systems to retrieve temporal information from texts about slots and their attribute relations.

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 2011 TAC KBP. The rest of the paper is organized in the following way: Section 2 introduces the source data and knowledge base used for the KBP track; Section 3 describes the training data for various 2011 KBP tasks. Section 4 discusses annotation and assessment procedures and methodologies; and Section 5 is the conclusion.

2 Source Data and Reference Knowledge Base

The monolingual tasks utilized the same document set as KBP 2010 (TAC 2010 KBP Source Data - LDC2010E12), which comprises 1.77 million

documents in several genres. New data selection was required for the cross-lingual task. LDC identified a set of 1 million newswire documents from Chinese Gigaword Fourth Edition (LDC2009T97), focusing on the 2007-2008 epoch to provide continuity with the English document collection (Table 1). Documents were drawn from three data providers: Xinhua News Agency, Agence France Presse and People's Daily Online. Documents with English translations appearing in LDC corpora were excluded. Document selection further considered coverage of entities also appearing in the English corpus.

The reference knowledge base (KB) (LDC2009E58) used for both the monolingual and cross-lingual Entity Linking tasks included 818,741 nodes – articles drawn from an October 2008 dump of English Wikipedia. Each node corresponds to a unique entity and to one of four types: person, organization, geo-graphical or unknown. 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.

Language	Genre	Documents
English	Broadcast Conversation	17
	Broadcast News	665
	Conversation Telephone Speech	1
	Newswire	1,286,609
	Web Text	490,596
Chinese	Newswire	1,000,000

Table 1: Document Source Collection

3 Training and Evaluation Data

For 2011, LDC created or revised training data for three tasks: Slot Filling, Temporal Slot Filling, and Cross-lingual Entity Linking. No new training data were developed for the monolingual Entity Linking task as past evaluation data was provided for this purpose. Evaluation data were created for the above three tasks as well as the English Entity Linking task. Table 2 lists the training and evaluation corpora for 2011 tasks.

Based on the results of the 2010 evaluation, significant revisions were made to the Slot Filling annotation guidelines, in particular the slot descriptions, in order to improve annotator consistency and ensure greater continuity between training data annotation on the one hand and assessment on the other hand. To ensure training and test data consistency, LDC then reviewed and updated assessments of fillers for the following slots from the 2009 evaluation and 2010 training and evaluation data sets: *ORG: Top Members/Employees, PER: Title, PER: Employee of, PER: Member of, ORG: Subsidiaries, PER: Nation of Residence, PER: State of Residence, PER: City of Residence, ORG: Members, ORG: Alternative Names, ORG: City of Residence, ORG: State of Headquarters, ORG: Nation of Headquarters*, and *ORG: Parents*. As a result of this corrections task, the assessments of 290 fillers (across all data sets) were changed from 'correct' to 'wrong'. Additionally, 18 fillers originally assessed as 'correct' were changed to 'inexact'.

For the Temporal Slot Filling task, LDC selected 40 identifiable PER entities and 10 identifiable ORG entities to serve as training queries, drawing from a pool of data developed under the DARPA Machine Reading program which had already been exhaustively annotated for KBP relations and their associated temporal information (the MR-TAC temporal superset, described further below). To adapt this data for use in TAC KBP, LDC converted the existing annotations into two output formats: an intermediate format containing the slot fillers plus temporal annotations in their original representation, and a new format containing the slot fillers plus temporal annotations represented as tuples [T1, T2, T3, T4]. While the tuple format functioned as the official Temporal Slot Filler format for TAC KBP, the intermediate format was included as a convenience for system developers.

After conversion of the training data to the official TAC KBP format, it became clear that a post-conversion screening process was necessary to make the training data fully conform to the TAC standards. First, fillers that did not match their slot value type had to be removed (e.g. *PER: Schools*

Corpus Title	Type	LDC Catalog	Language	Size (Queries)
TAC 2009 KBP Gold Standard Entity Linking Entity Type List	NEL Evaluation	LDC2009E86	English	567 GPE 627 PER 2710 ORG
TAC 2010 KBP Evaluation Entity Linking Gold Standard	NEL Evaluation	LDC2010E82	English	749 GPE 741 PER 750 ORG
TAC 2010 KBP Training Entity Linking	NEL Training	LDC2010E31	English	500 GPE 500 PER 500 ORG
TAC 2011 KBP Cross-lingual Training Entity Linking	NEL Training	LDC2011E55	Chinese English	685 GPE 817 PER 660 ORG
TAC 2010 KBP Training Slot Filling Annotation	SF Training	LDC2010E18	English	25 PER 25 ORG
TAC 2010 KBP Evaluation Slot Filling Annotation	SF Evaluation	LDC2010R11	English	50 PER 50 ORG
TAC 2011 KBP English Training Temporal Slot Filling Annotation	Temporal Training	LDC2011E49	English	40 PER 10 ORG
TAC 2011 KBP English Evaluation Entity Linking Annotation v1.1	NEL Evaluation	LDC2011R36	English	750 GPE 750 PER 750 ORG
TAC 2011 KBP Cross-lingual Evaluation Entity Linking Annotation V1.1	NEL Evaluation	LDC2011R38	Chinese English	642 GPE 824 PER 710 ORG
TAC 2011 KBP English Evaluation Regular Slot Filling Annotation V1.2	SF Evaluation	LDC2011R34	English	50 PER 50 ORG
TAC 2011 KBP English Evaluation Temporal Slot Filling Annotation	Temporal Evaluation	LDC2011R40	English	80 PER 20 ORG

Table 2: Training and Evaluation Data for Entity Linking and Slot Filling Tasks

attended is a name-value slot, so a string such as "racing school" is an invalid filler for that slot). Second, cross-document coreference of the slot fillers had to be carried out (e.g. "Bill" and "Bill Clinton") so that redundant fillers and tuples were not generated for a particular query-slot combination.

All training data for the Cross-lingual Entity Linking task was developed concurrently with the evaluation data in order to ensure consistency. A further description of this process is included below.

4 Annotation/Assessment Procedures and Methodologies

4.1 Entity Selection

Tagger outputs generated by an English name tagger (Grishman et al., 2005) on the English source collection and by a Chinese name tagger (Ji and Grishman., 2005) on the Chinese source collection were used as the basis for selecting entities for the Entity Linking and regular Slot Filling tasks. For any name string *n*, the number of documents containing *n* in the source data corpus

and the KB were counted in order to assess the likely productivity of an entity. Text strings generated by the named entity tagger were rejected if they did not meet the requirements for the EL or SF tasks or if they were incorrectly tagged, nonsensical, or included objectionable content.

From the tagger output, entities were selected for the EL tasks (monolingual and cross-lingual), based on their perceived level of confusability and diversity. Confusable entities were those with names ambiguous enough to refer to different entities. Selection preferences were given to entities with names that were confusable either within types (e.g. GPE "Newark" Delaware vs. "Newark" New Jersey), or across types (e.g. "Chicago" the city vs. "Chicago" the sports teams). Preference was also given to entities whose names had multiple renderings in the corpus (Table 3).

Name Diversity	**Examples**
Aliases	"Mark Twain" for "Samuel Clemens"
Nicknames	"Bobby", "Bob", "Rob", or "Bert" for "Robert"
Abbreviations	"UPenn" for "University of Pennsylvania"
Acronyms	"BBC" for British Broadcasting Corporation
Historical Forms	"Beiping" for "Beijing"
Honorifics	"Queen of England" for "Elizabeth"
Metaphor	"Iron Lady" for "Wu Yi of PRC"
Reordering	"Wang Fang" for "Fang Wang"
Misspelling	"Los Angoles" for "Los Angeles"
Deletion	"John Adams" for "John Quincy Adams"

Table 3: Name Diversity

Additionally, in order to increase confusability, entities were only marked as potentially appropriate for the EL task if they appeared to have either many (7+) or no possible KB node matches. In addition to confusability, entities were selected for EL based on their distribution along three dimensions: entity type, NIL versus non-NIL status, and genre. Selected entities were balanced by entity type, with 750 GPEs, 750 PERs, and 750 ORGs for a total of 2,250 queries. NIL versus non-NIL distribution was slightly skewed towards NILs. The distribution of genre was 2/3 newswire documents to 1/3 web documents for English entities, while all Chinese entities were drawn from newswire.

In the final stage of entity selection for EL, up to 20 reference documents containing the desired name string were selected from the source data through a GUI interface (Figure 1). If it was clear that the name string was used to refer to more than one entity in the corpus, the documents were selected based on their ability represent the different entities as equally as possible.

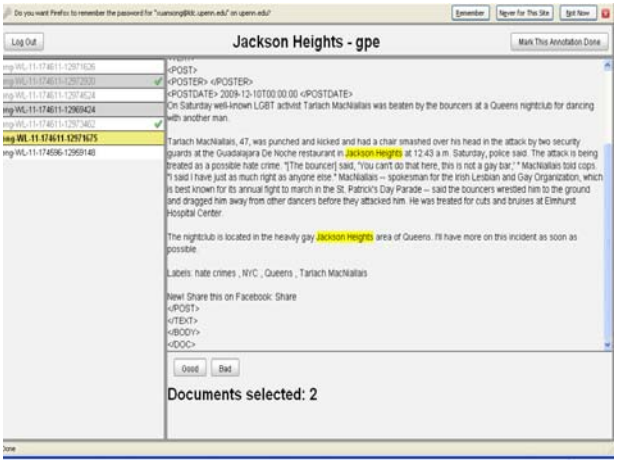

Figure 1: Reference Document Selection Tool

For the regular Slot Filling task, entities were selected from the tagger output based on their level of non-confusability and productivity. A candidate query was considered non-confusable if its name string was complete and was contained in at least one and no more than six KB entries. Productivity for candidate queries was determined by searching the source data corpus to determine whether it contained at least 2 - 3 slot fillers for the entity. During this search, reference documents were also selected for each entity.

Due to the sharing of temporally labeled data between TAC KBP and the Machine Reading program, the entity selection process for the Temporal Slot Filling task (TSF) utilized a different process than that used for regular Slot Filling. Rather than first selecting identifiable entities and then annotating slot fillers and

temporal information for those fillers, a reverse selection process was used in which annotation preceded entity selection.

The first step in this process was to perform keyword searches on the source data to identify documents containing KBP relations. The document set that resulted from this keyword search was then subsetted with a high keyword frequency threshold. The second step in this process was to screen this document subset for the presence of temporalized KBP relations. Resulting documents were then exhaustively annotated for KBP relations and their associated temporal information. This produced an annotation pool known as the MR-TAC temporal superset.

Because Machine Reading annotations do not require that the entities annotated in relations be identifiable (e.g. "she joined the company in 1981" is a valid relation instance in Machine Reading), a post-annotation screening process was necessary to select identifiable entities that could function as TSF queries. In this final screening process, annotators selected identifiable entities in the MR-TAC temporal superset that were part of at least one temporalized KBP relation; these entities then served as potential queries for the TSF.

Forty identifiable PER entities and 10 identifiable ORG entities were selected as TSF training queries. However, to ensure productivity of fillers and temporal information, an additional screening process was used to select the TSF evaluation queries. Evaluation query selection relied on the same post-annotation screening process to select identifiable candidate entities annotated in at least one temporalized KBP relation. Annotators then performed a time-limited search in the KBP source data for these candidate entities, to determine how frequently they occurred in temporalized KBP relations. Eighty identifiable PER entities and 20 identifiable ORG entities were then selected from the candidate evaluation query entity set, with preference given to entities that occurred more frequently in temporalized KBP relations and that occurred in a greater variety of temporalized KBP relations. This entity selection process produced a set of TSF evaluation queries on which time-limited search and cross-document annotation of temporalized slot fillers could be carried out during the evaluation annotation task.

4.2 Entity Linking and NIL-coreference

4.2.1 English Entity Linking

Reference documents selected for each entity in the monolingual Entity Linking task were reviewed by LDC annotators using a customized GUI developed by LDC for this task (Figure 2).

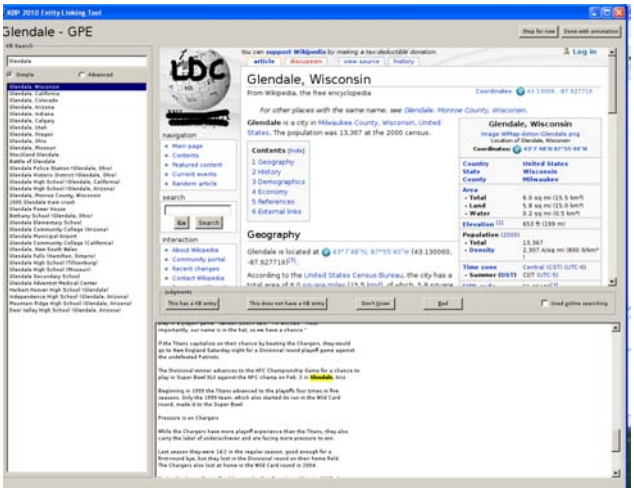

Figure 2: Entity Linking Annotation Tool

Annotators first reviewed the reference document then searched the KB for a matching node (2008 Wikipedia article), creating a link if a match was found. This annotation yielded two types of queries: those with KB matches (non-NILs) and those without KB matches (NILs). Non-NIL entities were automatically co-referenced when linked to the same KB node. NIL queries required manual co-reference annotation, which was performed in three separate passes (one per entity type). After the linking and coreference tasks were completed, a time-limited quality control pass was performed to enhance completeness and accuracy.

4.2.2 Cross-lingual Entity Linking

Cross-lingual Entity Linking was more complicated, both in entity selection and the annotation process, than the corresponding monolingual task. In addition to the entity selection criteria used for the monolingual task, Cross-Lingual Entity Linking required that the query selection be further balanced by language, with a targeted query distribution of 1/4 English-only, 1/4

Chinese-only, and 1/2 English-Chinese. The challenge was to select overlapping entities appearing in both Chinese and English source documents. To ease the selection process, LDC divided the tagger output into several bins from which confusable name strings were selected. First, by examining the number of KB nodes containing the name string, Chinese name lists were divided into NIL and non-NIL bins. Then using the number of English and Chinese source documents containing the name string, the NIL and non-NIL pools were further sorted into bins of English-only, Chinese-only and English-Chinese. The binning process resulted in a larger set of Chinese-only and English-only names, while the Chinese-English name set was very small, which explains the rarity of entities appearing both in Chinese and English source documents, especially for NIL entities.

To further augment the number of entities with mentions in both the English and Chinese corpora, LDC undertook an additional selection task in which bilingual annotators chose ambiguous names from the Chinese-English tagger output to create Chinese queries. They then translated the Chinese source names into English names to produce English queries denoting the same entities. Table 4 displays the language source distribution of the Cross-lingual Entity Linking training and evaluation datasets. "CMN" standards for "Chinese", and "ENG" for "English".

Datasets	CMN Only	ENG Only	CMN ENG	Total
2011 cross-lingual train	1399	461	302	2162
2011 cross-lingual eval	1408	548	220	2176

Table 4: Language Source Distribution for CLEL

The cross-lingual linking and co-reference annotation tasks were similar to their monolingual counterparts, but with two additions to co-reference annotation. To assure annotation quality, native English-speaking annotators co-referred the English query set while native Chinese-speaking annotators separately co-referred the Chinese sets. Afterward, bilingual annotators were tasked with co-referring the two datasets together. Before the final cross-lingual dataset was produced, a quality control pass was conducted on the linking and co-reference annotation, resulting in updates to approximately 10%-15% of the data.

4.3 Slot Filling

4.3.1 Annotation Approaches

Due to the need to prevent redundant annotations, Slot Filling in 2011 required a preliminary review of the KB nodes for all non-NIL entities selected for the task. During this review, slot fillers already contained in the KB were marked so that, upon being loaded into the annotation tool they would be visible to annotators but not editable.

As noted above, the Slot Filling annotation guidelines were significantly revised in 2011 following a review of data created in 2010. Potential 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 SF data. Only annotators who successfully completed testing were able to participate in the Slot Filling annotation task.

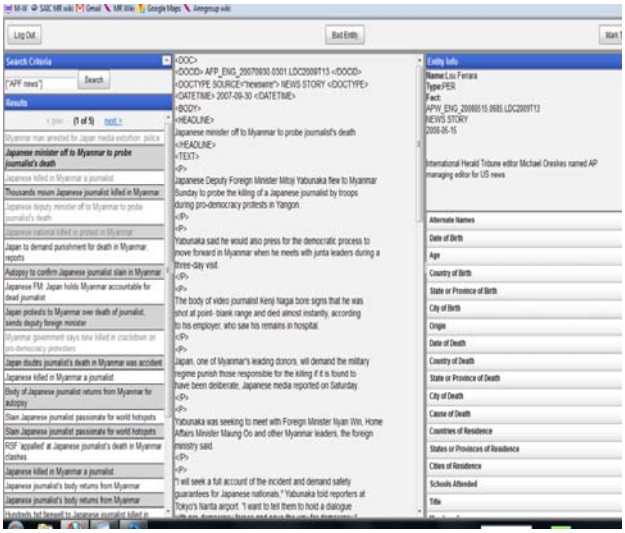

Figure 3: Slot Filling Annotation Tool

Annotation was performed using LDC's Slot Filling GUI (Figure 3 above), which includes a corpus search component and an annotation component. For each query entity annotators were given two hours in which to search the corpus and

locate fillers for the targeted slots. A quality control pass was conducted after Slot Filling annotation to flag any filler 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.

4.3.2 Slot Filling Assessment

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.

4.4 Temporal Slot Filling

4.4.1 Annotation Approaches

Temporal Slot Filling (TSF) was different from regular Slot Filling in that it sought to identify and capture temporal information in text that indicated *when* a slot and filler relation held true. Because this was a new task in this year's KBP evaluation, and because time and funding constraints prohibited the development of new annotation infrastructure, LDC made an effort to leverage existing temporal annotation tools from the DARPA Machine Reading MR KBP task. However, differences between the two programs' annotation requirements called for some workarounds to complete KBP TSF annotation.

The procedure had to account for the fact that, while TAC required annotators to search the entire corpus looking for temporal information relevant to any entity-slot-filler combination, MR KBP training annotations were completed on a document-by-document basis, which left the MR KBP annotation GUI with no search functionality. As a workaround, annotators used the search component from the existing TAC KBP slot filling GUI to conduct corpus searches, then used the annotation component from the Machine Reading MR KBP annotation GUI to perform temporal annotation, with some manual intervention required to port search results into the MR KBP tool.

The MR KBP temporal annotation format consists of labels (such as 'Beginning', 'Ending', or 'Within') on temporal expressions captured in text to indicate how the dates and durations are connected to the entity/filler relations. For example, given the entity 'Rudy Giuliani' and the sentence, "Rudy Giuliani lived in New York in 2000", an annotator would have connected "2000" to a residence relation between Mr. Giuliani and New York and labeled it as 'Within', meaning only that, at some point during the year 2000, the relation held true. Capturing information in this fashion posed something of a challenge to TSF annotators, who had been tasked with creating the fullest, text-supported 4-tuples to indicate when entity/filler relations held true, but were using an annotation tool that had not been designed to support the display of such tuples. As a result, temporal annotations were created and natively output in the MR KBP format and later converted into the official TAC tuple format as a post-processing step.

In order to reduce time taken up by the task, annotators were instructed not to annotate all temporal expressions related to entity/filler relations but only those required to support the fullest possible tuples. For example, if an annotator had already recorded that a relation began in 2006 and ended in 2008, there was no need to mark an expression indicating that the relation persisted in 2007.

4.4.2 Temporal Slot Filling Assessment

Assessment of TSF responses was divided into two tasks: assessment of slot fillers and assessment of temporal information connected to those fillers. The procedure used for assessing temporal slot fillers mirrored the process used for regular Slot Filling assessment.

After filler assessment was complete for the temporal data set, LDC compared the resulting list of documents containing correct, system-generated slot fillers with those annotated by humans during TSF. The purpose of this comparison was to identify all documents marked only by systems as containing temporal information for a given entity-slot-filler combination. Once these documents were identified, they were reviewed and annotated whenever temporal information relating to the specific entity-filler combination was present. Table 5 provides Human scores for TSF by slot:

Slot	F-Measure
PER: Countries of Residence	0.78
PER: State or Provinces of Residence	0.48
PER: Cities of Residence	0.28
PER: Employee of	0.74
PER: Member of	0.68
PER: Title	0.67
PER: Spouse	0.54
ORG: Top Member/Employee	0.55
Overall	0.66

Table 5: Human TSF Scores

5 Conclusion

This paper discussed procedures and methodologies for annotation and assessment for KBP 2011, particularly elaborating the challenges confronting the Cross-lingual Entity Linking task and the Temporal Slot Filling task. Future work will include additional focus on optimizing the entity selection process for overlapping Chinese/English queries for the Cross-lingual Entity Linking, as well as enhanced data selection methods to support Temporal Slot Filling. The resources described in this paper are slated for publication in the LDC Catalog, making 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.

Heng Ji and Ralph Grishman. 2005. Improving Name Tagging by Reference Resolution and Relation Detection. In Proceedings of ACL 05, Ann Arbor, USA.

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.