Entity Linking at TAC 2013

Task Description

Version 1.0 of April 9, 2013

1 Introduction

The main goal of the Knowledge Base Population (KBP) track at TAC 2013 is to promote research in and to evaluate the ability of automated systems to discover information about named entities and to incorporate this information in a knowledge source. For the evaluation an initial (or reference) knowledge base will be provided along with a source document collection from which systems are to learn. Attributes (a.k.a., "slots") derived from Wikipedia Infoboxes will be used to create the reference knowledge base. The overall task of populating a knowledge base is decomposed into two related tasks: Entity Linking (EL), where names must be aligned to entities; and Slot Filling (SF), which involves mining information about entities from text.

In the Entity Linking (EL) task, names must be aligned to entities in the reference KB and with other entities discovered in the collection. The task will be structured by having participants process a list of queries over target entities. This list will contain entity types of Person (PER), Organization (ORG), and Geo-Political Entity (GPE). As in the ACE evaluation, GPEs include inhabited locations with a government, such as cities and countries.

2 Entity Linking

2.1 Monolingual Entity Linking

In the Entity Linking task, given a query that consists of a name string, a background document ID, and a pair of UTF-8 character offsets indicating the beginning and end locations of the name string in the document, the system is required to provide the ID of the KB entry to which the name refers, or a "NILxxxx" ID if there is no such KB entry. The entity linking system is required to cluster queries referring to the same non-KB (NIL) entities and provide a unique ID for each cluster, in the form of NILxxxx (e.g., "NIL0021").

An example query is

```
<query id="EL_000304">
  <name>Robinson</name>
  <docid>APW_ENG_20080416.1425.LDC2009T13</docid>
  <beg>565</beg>
  <end>572</end>
</query>
```

The entity-linking system output file should contain one line for each query, where each line has three tab-delimited fields:

Field 1: query ID
Field 2: reference KB link (or NIL link)
Field 3: a confidence value

Each confidence value must be a positive real number between 0.0 (exclusive, representing the lowest confidence) and 1.0 (inclusive, representing the highest confidence), and must

include a decimal point (no commas, please). Up to five answers to a given query may be included in each submission. The main score for the task will use only the highest confidence answer for each query, selecting the answer that appears earliest in the submission if more than one answer has the highest confidence value. If practical given resource limitations, NIST will provide alternate scoring that looks at all submitted answers. At this point it is not feasible to allow multiple NIL clusterings; thus, no more than one NIL value may be submitted for a given query, and NIL clustering is not likely to play a part in the alternate scoring.

Entities will generally occur in multiple queries using different name variants and/or different docids. It is also expected that some entities will share confusable names (e.g., *George Washington* could refer to the president, the university, or the jazz musician; *Washington* could refer to a city, state, or person). For the primary task, the system may consult the text from the Wikipedia pages associated with the KB nodes. There will be also an optional task in which the systems should do linking without reference to these texts – using only the slot values; this corresponds to the task of updating a knowledge base with no 'backing' text.

2.2 Cross-lingual Entity Linking

The cross-lingual entity linking tasks follow monolingual entity linking; the steps are: (1) link non-NIL queries to English KB entries; and (2) cluster NIL queries. The cross-lingual aspect comes from the fact that the queries will include Chinese and Spanish queries. An example Chinese query is

```
<query id="EL_CMN_00007">
   <name>宜宾市</name>
   <docid>PDA_CMN_20070327.0068</docid>
   <beg>176</beg>
   <end>178</end>
</query>
```

2.3 Scoring Metric

For a set of query names with source documents, an entity linking system is required to: (1) judge whether each query can be linked to any KB node; and (2) partition all queries with NIL KB entries into clusters. Ultimately the system output can be viewed as a collection of clusters, some of which are labeled with KB node IDs. At the same time the answer key can also be viewed as a different collection of clusters over the same set. Therefore we will apply a modified B-Cubed metric (called B-Cubed+) to evaluate these clusters. Let us use the following notation:

$L(e)$ and $C(e)$ the category and the cluster of an item e,

$SI(e)$ and $GI(e)$ represent, respectively, the system (*i.e.,* participant-submitted) and gold-standard (ground truth) KB identifiers for an item e.

We define the correctness of the relation between e and e' in the distribution as:

$$G(e,e') = \begin{cases} 1 \, iff \; L(e) = L(e') \wedge C(e) = C(e') \wedge GI(e) = SI(e) = GI(e') = SI(e') \\ 0 \, otherwise \end{cases}$$

That is, two items are correctly related when they share a category if and only if they appear in the same cluster and share the same KB identifier in the system and the gold standard. B-cubed+ precision of an item is the proportion of correctly related items in its cluster (including itself). The

overall B-Cubed+ precision is the averaged precision of all items in the distribution. Since the average is calculated over items, it is not necessary to apply any weighting according to the size of clusters or categories. The B-Cubed+ recall is analogous, replacing "cluster" with "category". Formally:

$$\textit{Precision B-Cubed+} = Avg_e[Avg_{e'.C(e)=C(e')}[G(e,e')]]$$

$$\text{Recall } \textit{B-Cubed+} = Avg_e[Avg_{e'.L(e)=L(e')}[G(e,e')]]$$

The scorer will be available at: http://www.nist.gov/tac/2013/KBP/EntityLinking/tools.html

3 Data

3.1 Knowledge Base and Source Document Collection

The reference knowledge base includes hundreds of thousands of entities based on articles from an October 2008 dump of English Wikipedia, which includes 818,741 nodes.

Each entity in the KB includes the following:

- a name string;
- an assigned entity type of PER, ORG, GPE, or UKN (unknown);[1]
- a KB node ID (a unique identifier, like "E101");
- a set of 'raw' (Wikipedia) slot names and values; and
- some disambiguating text (*i.e.,* text from the Wikipedia page).

The 'raw' slot names and the values in the reference KB are based on the October 2008 Wikipedia snapshot. To facilitate use of the reference KB, a mapping from raw Wikipedia infobox slot-names to generic slots is provided in the training corpora.

The source document collection for the KBP 2013 Entity Linking tasks are composed of English, Spanish, and Chinese documents from the following LDC packages:

1. LDC2011T13: Chinese Gigaword Fifth Edition
2. LDC2011T07: English Gigaword Fifth Edition
3. LDC2011T12: Spanish Gigaword Third Edition
4. LDC2012E23: TAC 2012 KBP Source Corpus Additions Web Documents
5. Approximately half a million discussion forum posts in English, Chinese, and possibly Spanish (TBA)

Only a subset of the English, Spanish, and Chinese Gigaword collections will included in the official KBP 2013 tasks; the document IDs of those Gigaword documents that are part of the KBP 2013 tasks are listed in LDC catalog item LDC2012E22 (TAC 2012 KBP Source Corpus Additions Newswire Doc-ID Lists).

The following Table 1 presents the profile of the collection of source documents for the KBP 2013 entity-linking tasks.

Language	Source	Genre	Size (documents)
English	LDC2011T07	Newswire	1,000,257 (list in LDC2012E22)
	LDC2012E23	Web Text	999,999
	TBA	Discussion Fora	~500,000
Chinese	LDC2011T13	Newswire	2,000,256 (list in LDC2012E22)

[1] The reason this is abbreviated UKN rather than UNK is UKN.

	LDC2012E23	Web Text	815,886
	TBA	Discussion Fora	~500,000
Spanish	LDC2011T12	Newswire	1,000,020 (list in LDC2012E22)
	TBA	Discussion Fora	Up to ~500,000

Table 1. Distribution of Documents in KBP 2013 Source Document Collection

3.2 Training and Evaluation Corpus

The following tables summarize the KBP 2013 training and evaluation data that we aim to provide for participants. For all tasks we try to achieve a balance among genres, and between the queries with and without KB entry linkages.

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Training	2009 Eval	627	2710	567
	2010 Training Web data	500	500	500
	2010 Eval Newswire	500	500	500
	2010 Eval Web data	250	250	250
	2011 Eval Newswire	500	491	500
	2011 Eval Web data	250	259	250
	2012 Eval Newswire	702	388	381
	2012 Eval Web data	216	318	221
Evaluation (estimate)	2013 Newswire	333	333	333
	2013 Web/Discussion Fora data	333	333	333

Table 2. English Monolingual Entity Linking Data

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Training	2011 Training English/Chinese Newswire	250	250	250
	2011 Eval English/Chinese Newswire	250	250	250
	2012 Training Chinese Web	52	54	52

	2012 Eval Chinese Newswire	466	442	454
	2012 Eval Chinese Web	233	276	251
Evaluation (estimate)	2013 Eval Chinese Newswire	333	333	333
	2013 Eval Chinese Web/Discussion Fora	333	333	333

Table 3. Chinese Cross-lingual Entity Linking Data

Corpus	Genre/Source	Size (entity mentions)		
		Person	Organization	GPE
Training	2012 Training Spanish Newswire	664	597	563
	2012 Eval Spanish Newswire	669	539	858
Evaluation (estimate)	2013 Eval Spanish Newswire / Discussion Fora	667	667	667

Table 4. Spanish Cross-lingual Entity Linking Data

4 External Resource Restrictions and Sharing

4.1 External Resource Restrictions

As in KBP 2012, participants will be asked to make at least one run subject to certain resource constraints, primarily that the run be made as a 'closed' system … one which does not access the Web during the evaluation period. Sites may also submit additional runs that access the Web. This will provide a better understanding of the impact of external resources.

Further rules for both of the primary runs and additional runs are listed in Table 5.

Specific Rules	Specific Examples
Allowed	Using a Wikipedia derived resource to (manually or automatically) create training data
	Compiling lists of name variations based on hyperlinks and redirects prior to the evaluation
	Using a Wikipedia derived resource before evaluation to create a KB of world knowledge that can be used to check the correctness of facts
	Preprocess/annotate a large text corpus before the evaluation to check the correctness of facts or aliases
Not Allowed	Editing Wikipedia pages for target entities, either during, or after the evaluation

Table 5. Rules for Using External Resources

4.2 Resource Sharing

To support groups that intend to focus on part of the tasks, participants are encouraged to share external resources that they prepared before the evaluation. Such resources may include intermediate results, entity annotations, parsing/SRL/IE annotated Wikipedia corpus, topic model features for entity linking, etc.

5 Submissions

5.1 Submissions

In KBP 2013 participants will have one week after downloading the data to return their results for each task. Up to five alternative system runs may be submitted by each team for each task. Submitted runs should be ranked according to their expected score (based on development data, for example). Systems should not be modified once queries are downloaded. Details about submission procedures will be communicated to the track mailing list. The tools to validate formats will be made available at: http://www.nist.gov/tac/2013/KBP/EntityLinking/tools.html.

5.2 Schedule

Please visit the KBP 2013 entity-linking website for the schedule for the entity-linking tasks at KBP 2013: http://www.nist.gov/tac/2013/KBP/EntityLinking/

6 Mailing List and Website

The KBP 2013 website is http://www.nist.gov/tac/2013/KBP/. The website dedicated to the Entity-Linking tasks is http://www.nist.gov/tac/2013/KBP/EntityLinking/. Please post any questions and comments to the mailing list tac-kbp@nist.gov. Information about subscribing to the list is available at: http://www.nist.gov/tac/2013/KBP/registration.html.

Change History

- Version 1.0
  - Original version, based on the 2012 specification
  - Added confidence values and top N scoring