TAC KBP2014 Entity Linking Task Description

Version 1.1 of September 12, 2014

1 Overview

The Entity Linking (EL) track at NIST TAC-KBP2014 aims to extract named entity mentions from a source collection of textual documents, and link them to an existing Knowledge Base (KB). An EL system is also required to cluster mentions for those NIL entities that don't have corresponding KB entries. This year the task includes the following tracks: English (English text to English KB), Cross-lingual Chinese (Chinese and English text to English KB), and Cross-lingual Spanish (Spanish and English text to English KB).

Compared to the KBP entity linking evaluations in previous years, the main changes and improvement in KBP2014 include:

- Extend English task to Entity Discovery and Linking (full Entity Extraction + Entity Linking + NIL Clustering)
- Add discussion forums to Cross-lingual tracks (some entity morphs naturally exist in Chinese discussion forums)
- Provide basic NLP, full IE, Entity Linking and semantic annotations for some source documents to participants
- Increase ambiguity and variety in query selection to ensure that outstanding advanced algorithms that outperform naive baseline approaches can be identified
- Share some source collections and queries with regular and cold-start slot filling tracks

2 Mono-lingual English Entity Discovery and Linking

2.1 Motivations

Earlier entity linking work resolved one query at each time separately. Such non-collective approaches usually rely on prior popularity and context similarity with supervised models (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007; Milne and Witten, 2008b; Han and Sun, 2011). However, due to the lack of context, a collective inference model incorporating global evidence from deeper understanding of multiple entity mentions is more powerful than a non-collective approach that links each mention at a time. Recent work further leveraged the global topical coherence of all related mentions for collective inference by disambiguating a set of relevant mentions simultaneously (collective approaches), normally through supervised or graph-based reranking models (Cucerzan, 2007; Milne and Witten, 2008; Han and Zhao, 2009; Kulkarni et al., 2009; Pennacchiotti and Pantel, 2009; Ferragina and Sciatiella, 2010; Fernandez et al., 2010; Gottipati and Jiang, 2010; Radford et al., 2010; Cucerzan, 2011; Guo et al., 2011; Han and Sun, 2011; Han et al., 2011; Ratinov et al., 2011; Chen and Ji, 2011; Kozareva et al., 2011; Cassidy et al., 2012; Fahimi et al., 2012; Shen et al., 2013; Liu et al., 2013; Dalton and Dietz, 2013). Various coherence measures were introduced to choose the “collaborators” (related mentions), such as collaborative learning (Chen and Ji, 2011), ensemble ranking (Pennacchiotti and Pantel, 2009; Kozareva et al., 2011), co-occurred concept mentions (McNamee et al., 2011; Ratinov et al., 2011, Nguyen et al., 2012), topic modeling (Cassidy et al., 2012; Xu et al., 2012), relation extraction (Cheng and Roth, 2013), coreference (Nguyen et al., 2012; Huang et al., 2014), semantic relatedness (Nguyen et al., 2012; Chen et al., 2013; Huang et al., 2014), Neighborhood Expansion with Pseudo-Relevance Feedback (Dietz and Dalton, 2012), meta-paths (Huang et al., 2014 submission) and social networks (Cassidy et al., 2012; Huang et al., 2014). Furthermore, these two sub-problems - mention extraction and mention linking - can mutually enhance each other. Some recent work (Meij et al., 2012; Guo et al., 2013; Fahimi et al., 2013; Huang et al., 2014) jointly identify and disambiguate concept mentions.
2.2 Task Overview

Based on the above motivations, this year we will add a new task of Entity Discovery and Linking (EDL) in the mono-lingual English track. The goal is to conduct end-to-end entity extraction, linking and clustering. Given a document collection, an EDL system is required to automatically extract (identify and classify) entity mentions (“queries”), link them to the KB, and cluster NIL mentions (those that don’t have corresponding KB entries). Compared to Entity Linking from previous years, an EDL system needs to extract queries automatically. In contrast to Wikification (Bunescu and Pasca, 2006; Mihalcea and Csomai, 2007; Ratinov et al., 2011), EDL only focuses on three types of entities (PER, ORG and GPE) and requires NIL clustering. In order to evaluate the impact of entity name mention extraction on this new EDL task, we will also organize a diagnostic evaluation on English entity linking as defined in KBP2013, with perfect entity name mentions as input.

2.3 Input and Output

- **Input**
  The input to EDL is a set of raw documents. We will select a subset of the TAC 2014 document collection from multiple genres including newswire, web data, and discussion forum posts, which include high values in terms of both ambiguity and variety and substantial amount of NIL entity mentions.

- **Output**
  An EDL system is required to automatically generate the following two files.

  (1). Mention Query File
  An EDL system is required to identify and classify name mentions into person (PER), organization (ORG) or geo-political entity (GPE); and then represent each name mention as a query that consists of a name string, a document ID, and a pair of UTF-8 character offsets indicating the beginning and end locations of the name string in the document. The definition of offsets is the same as other tasks in KBP including slot filling. The detailed definition of an entity name mention (a query) is presented in the LDC query development guideline: [http://nlp.cs.rpi.edu/kbp/2014/elquery.pdf](http://nlp.cs.rpi.edu/kbp/2014/elquery.pdf).

  Each query entry will consist of the following five fields:

  - `<query id>` - A query ID, unique for each entity name mention.
  - `<name>` - The full namestring of the query entity mention.
  - `<docid>` - An ID for a document in the source corpus from which the namestring was extracted.
  - `<beg>` - The starting offset for the namestring.
  - `<end>` - The ending offset for the namestring.

  For example:

  `<query id="EL13_ENG_0001">`
  `<name>cairo</name>`
  `<docid>bolt-eng-DF-200-192451-5799099</docid>`
  `<beg>2450</beg>`
  `<end>2454</end>`
  `</query>`
(2). Link ID File
Then for each entity mention query, an EDL system should attempt to link it to the given knowledge base (KB). The EDL system is also 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”). It should generate a link ID file that consists of the entity type of the query, the ID of the KB entry to which the name refers, or a “NILxxxx” ID if there is no such KB entry. The link ID file should contain one line for each query, where each line has four tab-delimited fields:

Field 1: query ID
Field 2: reference KB link entity ID (or NIL link): A unique NIL ID or an entity node ID, correspondent to entity linking annotation and NIL-coreference (clustering) annotation respectively. If the entity node ID begins with "E", the text refers to an entity in the Knowledge Base (TAC 2009 KBP Evaluation Reference Knowledge Base). If the given query is not linked to an entity in the Knowledge Base (KB), then it is given a NIL-ID, which consists of "NIL" plus an integer (e.g. NIL001, NIL002). Both the entities with an entity node ID of "E" type and "NIL" type are assumed to be co-referenced (clustered), with the same "E" type ID or the same "NIL" ID if they refer to the same entity. Each "E" type ID and NIL ID is distinct from one another.
Field 3: entity type: \{GPE, ORG, PER\} type indicator for the entity
Field 4: 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.

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.

3 Cross-lingual Chinese/Spanish to English Entity Linking
3.1 Task
The cross-lingual entity linking tasks follow the monolingual entity linking in previous years in which the entity mention queries will be given; 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>

The main new change this year is adding discussion forum posts into both Spanish and Chinese source collections.
3.2 Entity Morphs in Chinese Discussion Forums

The information in traditional formal genres such as newswire is usually explicitly expressed. However, in some certain conditions users need to create new ways to communicate sensitive subjects in order to maximize communicative success and expressive power. For example, entity “morphs” widely exist in Chinese Twitter and Discussion forums (Huang et al., 2013). Morph is a special case of alias to hide the original objects (e.g., sensitive entities and events) for different purposes, including avoiding censorship (Bamman et al., 2012; Chen et al., 2013b), expressing strong sentiment, emotion or sarcasm, and making descriptions more vivid. Here is an example post using morphs:

“由于瓜爹的事情，方便面与天线摊牌. (Because of Gua Dad’s issue, Instant Noodles faces down with Antenna.)”, where

- “瓜爹(Gua Dad)” refers to “薄熙来(Bo Xilai)” because it shares one character “瓜(Gua)” with “薄瓜瓜(Bo Guagua)” who is the son of “薄熙来(Bo Xilai)”;
- “方便面(Instant Noodles)” refers to “周永康(Zhou Yongkang)” because it shares one character “康(kang)” with the well-known instant noodles brand “康师傅(Master Kang)”;
- “天线(Antenna)” refers to “温家宝(Wen Jiabao)” because it shares one character “宝(baby)” with the famous children’s television series “天线宝宝(Teletubbies)”;

Such entity morphs require an EL system to conduct deep linguistic analysis on the contexts because many of them cannot be linked to the real targets based on surface features.

4 Scoring Metric

The scorers are posted at http://nlp.cs.rpi.edu/kbp/2014/scoring.html

We use a tuple \((doc−id,start,end,entity−type,kb−id)\) to represent each entity mention, where a special type of kb-id is NIL. Let \(S = (s_1,\ldots,s_m)\) be the set of entity mentions in the system output, \(G = (g_1,\ldots,g_n)\) be the corresponding gold-standard. An output mention \(s_i\) matches a reference mention \(g_j\) iff:

1. \(s_i • doc−id = g_j • doc−id,\)
2. \(s_i • start = g_j • start, s_i • end = g_j • end,\)
3. \(s_i • entity−type = g_j • entity−type,\)
4. \(and s_i • kb−id = g_j • kb−id. \) (only considered in Linking performance)

We use \(C_g(s_i)\) to denote the cluster in \(G\) that contains the mention \(s_i\). If no mention in \(G\) can match \(s_i\), then \(C_g(s_i) = \emptyset\). Likewise, \(C_s(g_j)\) denotes the cluster in \(S\) that contains the mention \(g_j\).

We will report two scores: Linking performance, and Clustering performance. For Linking performance, we only consider non-NIL entity mentions, and report standard precision, recall, and F-measure on mention level with consideration of all of the above matching criteria:

\[
\text{Precision} = \frac{\# \text{ of correct non-NIL mentions in } S}{\# \text{ of non-NIL mentions in } S} \\
\text{Recall} = \frac{\# \text{ of correct non-NIL mentions in } S}{\# \text{ of non-NIL mentions in } G} \\
F_1 = \frac{2 \ast \text{Precision} \ast \text{Recall}}{(\text{Precision} + \text{Recall})}
\]
An example is as follows:

- Precision = (1+1)/6 = 1/3
- Recall = (1+1)/6 = 1/3
- F-score = 2 * 1/3 * 1/3 / (1/3 + 1/3) = 1/3

For clustering performance, we ignore kb-id and use the following two metrics to evaluate the accuracy.

### 4.1 B-Cube

We augment the B-Cubed metric to evaluate the entity linking NIL clusters. The recall of B-Cubed score is calculated as follows:

\[
\text{Recall} = \frac{\sum_{s_i \in S, C_g(s_i) \neq \emptyset} |C_g(s_i) \cap C_p(s_i)|}{|C_p(s_i)|}
\]

The numerator counts the ratio of each correct mention in its reference cluster, and the final recall is averaged by the total number of mentions in the gold-standard. Similarly, the precision of B-Cubed score is calculated as follows:

\[
\text{Precision} = \frac{\sum_{g_j \in G, C_p(g_j) \neq \emptyset} |C_p(g_j) \cap C_g(g_j)|}{|C_g(g_j)|}
\]

Finally the F₁ measure is reported as the final score:

\[
F_1 = F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

An example is illustrated below.
4.2 CEAF

In this metric, we first construct a one-to-one mapping between gold-standard clusters and system-output clusters. Then the precision, recall, and f-measure are computed based on each mapped pair of clusters. Let \( f \) be a one-to-one mapping for clusters in \( S \) and clusters in \( G \). Let \( \phi(C_s, C_g) \) be a non-negative similarity score of two clusters \( C_s \) and \( C_g \), and \( \phi(C_g, \emptyset) = 0 \). For instance, \( \phi(C_s, C_g) \) could be the number of matched entity mentions in the two clusters. The best mapping \( f^* \) can be found by:
\[ f^* = \arg\max_f \sum_{C_g} \phi (C_g, f(C_g)) \]

Luo 2005 formulated this problem as a maximal bipartite-matching problem, and showed that this can be solved by Kuhn-Munkres algorithm in polynomial time.

Based on the mapping, we calculate precision, recall, and F-measure as follows:

\[
\text{Precision} = \frac{\sum_{C_g} \phi (C_g, f^*(C_g))}{\sum_{C_g} f^*(C_g)}
\]

\[
\text{Recall} = \frac{\sum_{C_g} \phi (C_g, f^*(C_g))}{\sum_{C_g} f^*(C_g, C_g)}
\]

where the denominator are the self-similarity scores of gold-standard clusters and system-output clusters, respectively.

Finally, we define \( \phi \) to be the number of matched mentions in the two clusters:

\[ \phi(C_g, C_s) = |C_g \cap C_s| \]

An example is illustrated below.

CEAFm: Example

- **Solid**: best 1-1 alignment
- \( \phi(G_i, S_i) = |G_i \cap S_i| \)
- **Recall** = \#common / \#mentions-in-key = (2+1)/6 = 1/2
- **Precision** = \#common / \#mentions-in-response = (2+1)/6 = 1/2

We will also report many score variants by relaxing name mention boundary and type matching and removing the requirement of extracting nested mentions. We will also report diagnostic scores on name boundary identification F-measure, name classification accuracy and linking accuracy. For B-cubed and CEAF scores, we will report on different levels (entity level, mention level and document level). If time permits, we will also choose a subset of difficult queries that include substantial amount of ambiguity, variety and NILs. Then for these difficult queries, we will check: (1) The percentage of difficult queries
successfully discovered by the system; (2) Entity Linking Performance on the successfully discovered difficult queries

5 Data

5.1 Knowledge Base and Source Document Collection

The reference knowledge base includes 818,741 entity nodes based on articles from an October 2008 dump of English Wikipedia. To facilitate use of the reference KB, a partial mapping from raw Wikipedia infobox slot-names to generic slots is provided in the training corpora. 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 source document collection for the KBP 2014 Entity Linking tasks are composed of English, Spanish, and Chinese documents from the following LDC package: LDC2014E13 TAC 2014 KBP Source Corpus.

5.2 Training and Evaluation Corpora

The following tables summarize the KBP 2014 new training and evaluation data that we provide for participants. For all tasks we try to achieve a balance among genres. For cross-lingual entity linking tasks we also achieve a balance between the queries with and without KB entry linkages. The English EDL evaluation dataset will meet all of the same specifications as its training data counterpart.

---

\(^1\) The reason this is abbreviated UKN rather than UNK is UKN.
### Table 1. The Number of Entity Mentions in New Training Data for KBP2014 English Entity Linking and Discovery

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Genre</th>
<th>Entity Type</th>
<th>PER</th>
<th>GPE</th>
<th>ORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDC KBP2014 EDL Training Corpus (LDC2014E54)</td>
<td>Newswire</td>
<td>1992</td>
<td>2290</td>
<td>6681</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussion Forum</td>
<td>285</td>
<td>192</td>
<td>230</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2277</td>
<td>2482</td>
<td>6911</td>
<td></td>
</tr>
<tr>
<td>KBP2014 AMR EDL Corpus Prepared by RPI and ISI (LDC2014E15)</td>
<td>Newswire</td>
<td>1178</td>
<td>776</td>
<td>819</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Web data</td>
<td>539</td>
<td>294</td>
<td>331</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Discussion Forum</td>
<td>1476</td>
<td>270</td>
<td>283</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>3193</td>
<td>1340</td>
<td>1433</td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. English Monolingual Entity Linking Data from 2009-2013

<table>
<thead>
<tr>
<th>Genre/Source</th>
<th>Size (entity mentions)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Person</td>
</tr>
<tr>
<td>2009 Eval</td>
<td>627</td>
</tr>
<tr>
<td>2010 Training Web data</td>
<td>500</td>
</tr>
<tr>
<td>2010 Eval Newswire</td>
<td>500</td>
</tr>
<tr>
<td>2010 Eval Web data</td>
<td>250</td>
</tr>
<tr>
<td>2011 Eval Newswire</td>
<td>500</td>
</tr>
<tr>
<td>2011 Eval Web data</td>
<td>250</td>
</tr>
<tr>
<td>2012 Eval Newswire</td>
<td>702</td>
</tr>
<tr>
<td>2012 Eval Web data</td>
<td>216</td>
</tr>
<tr>
<td>2013 Eval Newswire</td>
<td>333</td>
</tr>
<tr>
<td>2013 Eval Web/Discussion Fora data</td>
<td>333</td>
</tr>
<tr>
<td>Corpus</td>
<td>Genre/Source</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Training</td>
<td>2011 Training English/Chinese Newswire</td>
</tr>
<tr>
<td></td>
<td>2011 Eval English/Chinese Newswire</td>
</tr>
<tr>
<td></td>
<td>2012 Training Chinese Web</td>
</tr>
<tr>
<td></td>
<td>2012 Eval Chinese Newswire</td>
</tr>
<tr>
<td></td>
<td>2012 Eval Chinese Web</td>
</tr>
<tr>
<td></td>
<td>2013 Eval Chinese Newswire</td>
</tr>
<tr>
<td></td>
<td>2013 Eval Chinese Web</td>
</tr>
<tr>
<td></td>
<td>2014 Training Chinese Discussion Fora</td>
</tr>
<tr>
<td>Evaluation</td>
<td>2014 Eval Chinese Newswire</td>
</tr>
<tr>
<td>(estimate)</td>
<td>2014 Eval Chinese Web</td>
</tr>
<tr>
<td></td>
<td>2014 Eval Chinese Discussion Fora</td>
</tr>
</tbody>
</table>

Table 3. Chinese Cross-lingual Entity Linking Data

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Genre/Source</th>
<th>Size (entity mentions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>2012 Training Spanish Newswire</td>
<td>664 597 563</td>
</tr>
<tr>
<td></td>
<td>2012 Eval Spanish Newswire</td>
<td>669 539 858</td>
</tr>
<tr>
<td></td>
<td>2013 Eval Spanish Newswire / Discussion Fora</td>
<td>667 667 667</td>
</tr>
<tr>
<td></td>
<td>2014 Training Spanish Discussion Fora</td>
<td>166 166 166</td>
</tr>
<tr>
<td>Evaluation</td>
<td>2014 Eval Spanish Newswire</td>
<td>222 222 222</td>
</tr>
<tr>
<td>(estimate)</td>
<td>2014 Eval Spanish Web</td>
<td>222 222 222</td>
</tr>
<tr>
<td></td>
<td>2014 Eval Spanish discussion Fora</td>
<td>222 222 222</td>
</tr>
</tbody>
</table>

Table 4. Spanish Cross-lingual Entity Linking Data

### 6 Closed Test and Open Test

Participants will be asked to make at least one run subject to only using the resources and data sets provided by KBP2014 for the core linking component (KB ID candidate retrieval and ranking) as official results. No online web search is allowed for the official run. Sites may also submit additional resources, which will be also reported for comparison.

#### 6.1 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 entity linking systems, intermediate results, entity annotations, parsing/SRL/IE annotated Wikipedia corpus, topic model features for entity linking, etc. A recommended reading list of papers is at [http://nlp.cs.rpi.edu/kbp/2014/elreading.html](http://nlp.cs.rpi.edu/kbp/2014/elreading.html) and a list of publicly available softwares is at: [http://nlp.cs.rpi.edu/kbp/2014/tools.html](http://nlp.cs.rpi.edu/kbp/2014/tools.html). The participants are also encouraged to participate in other related tasks: [http://nlp.cs.rpi.edu/kbp/2014/events.html](http://nlp.cs.rpi.edu/kbp/2014/events.html) Please contact Heng Ji <jih@rpi.edu> if you want to share any resources through the KBP2014 website.

### 7 Submissions

In KBP 2014 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://nlp.cs.rpi.edu/kbp/2014/tools.html

8 Schedule

- February 21: Release Abstract Meaning Representation (AMR) Entity Linking Corpus
- March 28: New English Entity Linking scorer B-cubed++ available
- April 1: Release Track specification
- April 15: Release source collection
- May 23: Release English Entity Discovery and Linking training data
- May 23: Release Chinese EL Training Data
- May 23: Release Spanish EL Training Data
- June 15: Deadline for registration for track participation
- September 15-September 22: Cross-lingual Chinese-English Entity Linking Evaluation window
- September 15-September 22: Cross-lingual Spanish-English Entity Linking Evaluation window
- September 22 9am EDT-September 30 7am EDT: Mono-lingual English Entity Discovery and Linking evaluation window
- September 30 9am EDT - October 3 9am EDT: Mono-lingual English EDL Diagnostic Evaluation Window (Perfect Entity Mentions are given)
- September 25: Release Cross-lingual Entity Linking evaluation results
- Oct 2: Release Mono-lingual English EDL evaluation results
- October 7: Participants short system description due at NIST (for coordinators' overview paper)
- October 7: Presentation proposals due for all tracks. Teams are strongly encouraged to submit initial proposals based on system approaches and results on DEV data (proposals may be revised or withdrawn after NIST returns evaluation results)
- October 22: Notification of acceptance of presentation proposals
- November 1: Participants' full workshop papers due at NIST
- November 17-18: TAC KBP 2014 Workshop

- February 15, 2015: System description paper camera ready

9 Mailing List and Website

The KBP 2014 Entity Linking website is http://nlp.cs.rpi.edu/kbp/2014/. Please post any questions and comments to the mailing list tac-kbp@nist.gov. Information about subscribing to the list is available at: http://nlp.cs.rpi.edu/kbp/2014/mailing.html.

10 References and Tools

http://nlp.cs.rpi.edu/kbp/2014/elreading.html
http://nlp.cs.rpi.edu/kbp/2014/tools.html