TAC KBP2015 Entity Discovery and Linking Task Description

Version 0.4 of February 16, 2015

1 Overview
The Tri-lingual Entity Discovery and Linking (EDL) track at NIST TAC-KBP2014 aims to extract named entity mentions from a source collection of textual documents in multiple languages (English, Chinese and Spanish), and link them to an existing Knowledge Base (KB). An EDL system is also required to cluster mentions for those NIL entities that don't have corresponding KB entries. Compared to the KBP2014 EDL task, the main changes and improvement in KBP2015 include:
- Extend English EDL task from mono-lingual to tri-lingual
- A new KB: we will use a Freebase snapshot
- Add two new entity types - natural locations (LOC) and facilities (FAC)
- A new diagnostic task of English Entity Discovery within the Cold-start KBP track

2 Task Definition

2.1 Motivations
We have achieved some promising successes in English Entity Discovery and Linking in the previous years. However, for certain entities, a lot of information is only available in documents written in a foreign language for which there may be very few linguistic resources (annotated data, tools, etc.) available. In order to promote the new research direction on cross-lingual knowledge transfer to improve the portability of EDL techniques, this year we extend EDL from mono-lingual to cross-lingual.

2.2 Task Overview
Given a document collection in three languages (English, Chinese and Spanish), a tri-lingual EDL system is required to automatically identify, classify, cluster and link entity mentions to the English KB, and cluster NIL mentions (those that don’t have corresponding KB entries). The general skeleton of the task remains the same as English EDL, but the entities should be discovered from documents in three languages instead of one.

2.3 Input and Output

- Input
The input to EDL is a set of raw documents in English, Chinese and Spanish.

- Output
An EDL system is required to identify and classify name mentions into person (PER), organization (ORG), geo-political entity (GPE), location (LOC) and facility (FAC); 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 and entity mention annotation guidelines: http://nlp.cs.rpi.edu/kbp/2015/annotation.html. Then for each entity mention, 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.
An EDL system is required to automatically generate an output file, which contains one line for each mention, where each line has the following tab-delimited fields:

Field 1: system run ID
Field 2: mention (query) ID: unique for each entity name mention.
Field 3: name string: the full name string of the query entity mention.
Field 4: document ID: name start offset - name end offset: an ID for a document in the source corpus from which the name string was extracted, the starting offset of the name string, and the ending offset of the names string.
Field 5: 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. 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 6: entity type: {GPE, ORG, PER, LOC, FAC} type indicator for the entity
Field 7: 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.

2.4 Diagnostic Tasks

Teams can also submit results from 1 language or 2 languages. We will also setup a diagnostic Entity Linking task by providing systems perfect entity mentions. The output format of this diagnostic Entity Linking task is the same as the full EDL task. The input is a query file including five fields:

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

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>
```

There will be also a diagnostic task of Entity Discovery in the cold-start KBP track, which will be described in the cold-start task specification.
3 Scoring Metric

We will use various scoring metrics from KBP2014 EDL task. The detailed description is in section 2.1.2 in the overview paper: http://nlp.cs.rpi.edu/paper/edl2014overview.pdf. The scorers are posted at http://nlp.cs.rpi.edu/kbp/2015/scoring.html

4 Data

The reference knowledge base is a January 2015 snapshot of English Freebase. The source collection will approximately include 500 training documents and 500 evaluation documents. Some mentions will be from single language only, and some mentions will exist in multiple languages to form cross-lingual entity clusters. The corpus will consist of topic-focused news articles and discussion forum posts published in recent years, comparable (but non-parallel) across languages. The discussion forum posts may include code-switching in multiple languages.

5 Resources

No online web search is allowed for the official run. We encourage participants to investigate novel research directions beyond “baseline frameworks” (e.g., full-document machine translation + English EDL; Chinese/Spanish EDL linking to Chinese/Spanish KB + linking to English KB with explicit, manually created inter-lingual KB links). To support groups that intend to focus on part of the tasks, participants are encouraged to share external resources and tools that they prepared before the evaluation. A recommended reading list of papers is at http://nlp.cs.rpi.edu/kbp/2015/elreading.html and a list of publicly available softwares is at: http://nlp.cs.rpi.edu/kbp/2015/tools.html. The participants are also encouraged to participate in other related tasks: http://nlp.cs.rpi.edu/kbp/2014/events.html.

6 Submissions

In KBP 2015 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/2015/tools.html

7 Schedule (Tentative)

- January 17: Annotation Guidelines available
- February 16: Release task spec initial version
- March 1: Release Final version of Task spec
- March 15: Make previous years’ training/eval data sets available to participants
- March 15: new KB ready
- March 15: release source collection for Pilot Study
- March 15: Tri-lingual EDL related tools and resources available
- March 30: Pilot Registration deadline
- April: Pilot Training Data available
- May 1: Pilot Evaluation
- May 15: Full Training Data available
- September 1: Registration deadline
- Oct 5-12: Evaluation
- October 19: Participants short system description due at NIST (for coordinators' overview paper)
• October 19: 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: Coordinator's overview paper & Participants' full workshop papers due at NIST
• November 17-18: TAC KBP 2015 Workshop
• February 15, 2016: System description paper camera ready

8 Mailing List and Website

The KBP 2015 Entity Discovery and Linking website is http://nlp.cs.rpi.edu/kbp/2015/. 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/2015/mailing.html.

9 References and Tools

http://nlp.cs.rpi.edu/kbp/2015/elreading.html
http://nlp.cs.rpi.edu/kbp/2015/tools.html