TAC KBP2016 Entity Discovery and Linking Task Description

Version 1.0 of March 9, 2016

1 Overview

The Tri-lingual Entity Discovery and Linking (EDL) track at NIST TAC-KBP2016 aims to extract 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 KBP2015 EDL task, the main changes and improvement in KBP2016 include:

- Combine with tri-lingual slot filling to form up an end-to-end tri-lingual KBP task.
- Target at a larger scale data processing, by increasing the size of source collections from 500 documents to 90,000 documents. 500 "core" documents will be manually annotated for evaluation.
- Specific, individual nominal mentions will be expanded to all entity types and all languages.

2 Task Definition

2.1 Major Change: Connection with Tri-lingual Slot Filling

In KBP2016, the major change on EDL is to connect it with Tri-lingual slot filling, so an end-to-end Tri-lingual KBP pipeline will be complete. Therefore, we will be using the same source collection as Tri-lingual cold-start slot filling, which approximately consists of 90,000 documents. 500 documents will be selected for EDL evaluation. LDC will perform manual EDL annotations for these 500 documents. Because of the dramatic increase of data size, we will have two evaluation windows:

- **Aug 1-10: Tri-lingual EDL First Evaluation Window (EDL1)**
- **Sept 12-19: EDL Second Evaluation window (EDL2)**

The results of EDL1 systems will be used by Tri-lingual slot filling systems, therefore it is a great opportunity to measure the impact of EDL on end-to-end KBP. EDL2 systems can take advantage of output from EDL1 and slot filling results. An EDL team is highly encouraged to participate in both EDL1 and EDL2.

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 entity mentions into pre-defined entity types: Person (PER), Geo-political Entity (GPE), Organization (ORG), Location (LOC), Facility (FAC). In 2016, entity mentions include both name mentions and nominal mentions for all entity types and all three languages. Nominal mentions are limited to specific and individual mentions.

The definition of offsets is the same as other tasks in KBP including slot filling. The detailed definition of an entity mention (a query) is presented in the LDC query development guideline and entity mention annotation guidelines: http://nlp.cs.rpi.edu/kbp/2016/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:

Filed 1: system run ID
Field 2: mention (query) ID: unique for each entity name mention.
Field 3: mention head string: the full head string of the query entity mention.
Field 4: document ID: mention head start offset – mention head end offset: an ID for a document in the source corpus from which the mention head was extracted, the starting offset of the mention head, and the ending offset of the mention head.
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.
Field 6: entity type: {GPE, ORG, PER, LOC, FAC} type indicator for the entity
Field 7: mention type: {NAM, NOM} type indicator for the entity mention
Field 8: 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.

- **Offset Calculation and Formatting**

Each document is represented as a UTF-8 character array and begins with the "<DOC>" tag, where the "<" character has index 0 for the document. Thus, offsets are counted before XML tags are removed. The start offset must be the index of the first character in the corresponding string, and end offset must be the index of the last character of the string (therefore, the length of the corresponding mention string is endoffset – startoffset + 1). Start and end offsets should be separated by a dash ("-") with no surrounding spaces.

2.4 Diagnostic Task

Teams can also submit EDL results from only 1 language or 2 languages. The output format is the same as the full EDL task. We will report diagnostic scores. In 2016 we don't have a separate Entity Linking diagnostic task that uses perfect entity mentions.

3 Scoring Metric

We will apply various scoring metrics from KBP2015 EDL task. The detailed description is in section 2.2 in the overview paper: http://nlp.cs.rpi.edu/paper/edl2015overview.pdf. The scorers are posted at http://nlp.cs.rpi.edu/kbp/2016/scoring.html

4 Data

Same as last year, the reference knowledge base is BaseKB, which is available from LDC as LDC2015E42: TAC KBP 2015 Tri-Lingual Entity Discovery and Linking Knowledge Base. To evaluate

the scalability of EDL systems, and better connect with slot filling systems, the source collection will include 90,000 documents in 2016. 500 documents will be selected for evaluation. LDC will prepare manual annotations for those 500 documents. The following changes will be made on annotation in 2016:

- Specific, individual nominal mentions are added for all entity types and all languages.
- Titles will not be separately annotated
- Embedded mentions (those within a token) will be discontinued

5 Submissions

During each evaluation window, 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). No online web search is allowed for the official run. Systems should not be modified once data sets 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/2016/tools.html

6 Resources

The available data sets are summarized in the evaluation license: http://www.nist.gov/tac/2016/KBP/TACKBP16_eval_license_V1.0.pdf 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.

Some recommended reading lists of papers are at http://nlp.cs.rpi.edu/kbp/2016/elreading.html and http://nlp.cs.rpi.edu/kbp/2016/sfreading.html

A list of publicly available softwares is at: http://nlp.cs.rpi.edu/kbp/2016/tools.html

7 Schedule (Tentative)

- March 7: Tri-lingual EDL related tools and resources available
- March 9: Release task spec initial version
- March 20: Release Final version of Task spec
- March 31: Make previous years' training/eval data sets available to participants
- April 30: Possible dry run of cross-lingual Spanish-English SF task
- July 15: Registration deadline
- **Aug 1-10: Tri-lingual EDL First Evaluation Window (EDL1) on 90K docs**
- **Aug 15-29: Tri-lingual Slot Filling Evaluation Window on 90K docs (optional input/resource: EDL1 runs)**
- Sept 1: Release EDL1 scores to individual participants
- **Sept 12-19: EDL Second Evaluation window (EDL2) on 90K docs (optional input/resource: CSSF, CSKB runs, EDL Assembling Run)**
- Sept 25: Release EDL2 scores to individual participants
- October 10: Participants short system description due at NIST (for coordinators' overview paper)
- October 10: Presentation proposals due for all tracks
- October 10: Notification of acceptance of presentation proposals
- November 1: Coordinator's overview paper & Participants' full workshop papers due at NIST
- November 14-15: TAC KBP 2016 Workshop
- February 15, 2017: System description paper camera ready

8 Pilot Studies

In order to promote interesting research methods and identify future directions of KBP, participants are welcome to join RPI team to perform the following informal pilot studies. Please contact the coordinator Heng Ji (jih@rpi.edu) if interested:

- New Entity types: Add "named classes", Weapon/Vehicle, or more fine-grained entity types (RPI has cleaned up 9000+ types based on YAGO), or allow EDL systems to automatically discover new entity types;
- EDL on streaming data
- Surprise low-resource language EDL and slot filling

9 Mailing List and Website

The KBP 2016 Entity Discovery and Linking website is http://nlp.cs.rpi.edu/kbp/2016. 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/2016/mailing.html.

10 Organizing Committee

Hoa Trang Dang (U.S. National Institute of Standards and Techonology, hoa.dang@nist.gov)
Jason Duncan (MITRE, jduncan@mitre.org)
Joe Ellis (Linguistic Data Consortium, joellis@ldc.upenn.edu)
Marjorie Freedman (BBN Technologies, mfreedman@bbn.com)
Jeremy Getman (Linguistic Data Consortium, jgetman@ldc.upenn.edu)
Ralph Grishman (New York University, grishman@cs.nyu.edu)
Ben Hachey (University of Sydney, ben.hachey@sydney.edu.au)
Heng Ji (Coordinator, Rensselaer Polytechnic Institute, jih@rpi.edu)
Joel Nothman (University of Sydney, joel@it.usyd.edu.au)
James Mayfield (Johns Hopkins University, james.mayfield@jhuapl.edu)
Boyan Onyshkevych (U.S. Department of Defense, boyan.onyshkevych@darpa.mil)
Zhiyi Song (Linguistic Data Consortium, zhiyi@ldc.upenn.edu)
Stephanie Strassel (Linguistic Data Consortium, strassel@ldc.upenn.edu)