TAC KBP2017 Entity Discovery and Linking Pilot
on 10 Low-Resource Languages

Version 1.0 of April 26, 2017

1 Overview

Research on the TAC-KBP Entity Discovery and Linking (EDL) task has made good progress in the past three years, especially on foreign languages like Chinese and Spanish. In EDL2016, the top Chinese and Spanish systems achieved comparable performance as the top English systems [1]. However, most of the success was due to clean manual annotation efforts made by LDC or participants. Clean data annotation is often not available for low-resource languages and difficult to obtain during emergent settings. In order to compensate this data requirement, various automatic annotation generation methods have been proposed to create "Silver Standard", including knowledge base driven distant supervision, cross-lingual projection, and leveraging naturally existing noisy annotations such as Wikipedia markups [2]. In this pilot we will perform cross-lingual name tagging and linking for ten low-resource languages, and aim to answer the following research questions:

- How to fill in the performance gap between silver standard and gold standard?
- Can we advance the field by exploring non-traditional linguistic resources which are beyond human data annotation?
- Is there any performance ceiling for cross-lingual EDL? To what extent is it due to the lack of language-specific knowledge?
- Silver-standard annotations are usually very noisy, while many machine learning methods are sensitive to noise. How to make these learning models more robust to noise?

2 Task Definition

2.1 Foreign Language Choice

NIST and DARPA have chosen the following ten low-resource languages to compose our source collection, by considering multiple factors including the amount of available resources, language diversity and end user needs:

- Polish
- Chechen
- Albanian
- Swahili
- Kannada
- Yoruba
- Northern Sotho
- Nepali
- Kikuyu
- Somali

2.2 Task Overview

Given a document collection in any of the above ten languages, a cross-lingual EDL system in this pilot study is required to automatically identify and classify named entity mentions into four pre-defined entity types: Person (PER), Geo-political Entity (GPE), Organization (ORG) and Location (LOC), and link each mention to an English KB (a Wikipedia dump will be provided) if it is linkable, and otherwise produce "NIL" if it's not linkable.

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" label or an entity node ID, correspondent to entity linking annotation.
Field 6: entity type: {GPE, ORG, PER, LOC} type indicator for the entity
Field 7: mention type: NAM
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.

3 Scoring Metric

We will apply various scoring metrics from KBP2016 EDL task, and use name tagging F-score and entity linking accuracy as the main metrics for comparison. The detailed description is in section 2.2 in the overview paper: http://nlp.cs.rpi.edu/paper/edl2016overview.pdf. The scorers are posted at http://nlp.cs.rpi.edu/kbp/2016/scoring.html

4 Data

Evaluation Data: RPI and LDC will provide gold-standard annotations for 1,000-4,000 name mentions for each language.

Training Data: RPI and other LORELEI sites will provide silver-standard annotations, derived from various methods, covering 100 ~ 1 million name mentions for each language.

Non-Traditional Linguistic Resources: RPI will provide cleaned non-traditional linguistic resources for some languages, including those derived from WALS, CIA Names, Grammar Book, PanLex, Wikitionaries, and Language Survival Kits.

Other Resources provided by LDC: For 3 Languages in LORELEI LRLPs, LDC will also provide monolingual data, comparable data, parallel data, lexicon, etc.

5 Submissions

During each evaluation window, participants will have 12 days 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 language. 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/2017/tools.html

6 Resources

Teams are encouraged to exploit any non-traditional linguistic resources and features. RPI will also provide baseline systems for all of these languages.

7 Schedule

- April 28: Release task spec
- July 15: Registration deadline
- August 25: Silver-Standard Training Data available
- **October 5-16: Evaluation Window**
- October 18: Release scores to individual participants
- October 22: Participants short system description due at NIST (for coordinators' overview paper)
- October 22: Presentation proposals due for all tracks
- October 23: Notification of acceptance of presentation proposals
- November 1: Coordinator's overview paper & Participants' full workshop papers due at NIST
- November 13-14: TAC KBP 2016 Workshop
- February 15, 2018: System description paper camera ready

8 Mailing List and Website

The KBP 2017 Entity Discovery and Linking website is http://nlp.cs.rpi.edu/kbp/2017. 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/2017/mailing.html.

9 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)
Jeremy Getman (Linguistic Data Consortium, jgetman@ldc.upenn.edu)
Heng Ji (Coordinator, Rensselaer Polytechnic Institute, jih@rpi.edu)
Joel Nothman (University of Sydney, joel@it.usyd.edu.au)
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)

References

[1] Heng Ji, Joel Nothman and Hoa Trang Dang. 2016. Overview of TAC-KBP2016 Tri-lingual EDL and Its Impact on End-to-End KBP. Proc. Text Analysis Conference (TAC2016).

[2] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight and Heng Ji. 2017. Cross-lingual Name Tagging and Linking for 282 Languages. Proc. the 55th Annual Meeting of the Association for Computational Linguistics (ACL2017).