TAC KBP2018 Entity Discovery and Linking for Thousands of Entity Types

Version 0.1 of May 1, 2018

1 Motivations and Goals

The goal of TAC-KBP Entity Discovery and Linking (EDL) is to extract mentions of pre-defined entity types from any language, and link (disambiguate and ground) them to the entities in an English knowledge base (KB). In the past several years we have only focused on five major coarse-grained entity types: person (PER), geo-political entity (GPE), location (LOC), organization (ORG) and facility (FAC). Many real world applications in scenarios such as disaster relief and technical support require us to significantly extend our EDL capabilities to a wider variety of entity types (e.g., technical terms, lawsuits, disease, crisis, vehicles, food, biomedical entities). Table 1 presents some examples of the 16K+ entity types defined in YAGO (Suchanek et al., 2007; Mahdisoltani et al., 2015).

Language	Example
English	∗ *Briain-derived neuotrophic factor*$_{\texttt{Hormone}}$, another important gene in neural plasticity .
	∗ He was reputed to be implicated in the *Popish Plot*$_{\texttt{Fraud}}$.
	∗ *Brown v. Board of Education*$_{\texttt{Lawsuit}}$ was a landmark *United States Supreme Court*$_{\texttt{Assembly}}$ case .
	∗ She carried one *15 cm SK L/45 gun*$_{\texttt{NavalGun}}$, and one *8.8 cm SK L/30 gun*$_{\texttt{NavalGun}}$ in a U - boat mounting .
Italian	∗ *Pleoticus muelleri*$_{\texttt{Seafood}}$, nome comune *Gambero argentino*$_{\texttt{Seafood}}$ è un crostaceo decapodo .
	(*Translation: Pleoticus muelleri*$_{\texttt{Seafood}}$, common name *Argentine shrimp*$_{\texttt{Seafood}}$ is a decapod crustacean.)

Table 1. Some sentence examples annotated with YAGO entity types

In TAC-KBP2018 we will extend the number of types from five to thousands defined in YAGO. In addition, we will ask human analysts to provide "feedback" on system output, which will be used to improve each system.

2 Task Overview

Given a document collection in English, an EDL system is required to automatically identify and classify entity name mentions into one of the types defined in the schema (section 3), and link each mention to an English KB if it is linkable, and otherwise produce "NIL" if it's not linkable. The 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 entity name 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: a 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.

The tools to validate formats will be made available at: http://nlp.cs.rpi.edu/kbp/2018/tools.html

3 Entity Ontology/Schema

We selected 7,309 entity types from YAGO/WordNet. Each type has at least 10 entity entries in DBPedia. The se entity types are at the following link (in JSON format):
https://nlp.cs.rpi.edu/kbp/2018/yago_types_of_at_least_10.json
key: YAGO/WordNet type; value: a list of entities which have this type.

The visualized schema is at:
https://blender04.cs.rpi.edu/~panx2/tmp/typing/taxonomy/

Some examples of English and Chinese Wikipedia sentences including these types are at:
https://blender04.cs.rpi.edu/~panx2/tmp/typing/

4 Data Sets

- **Source Collection and Annotation**

The following source document collections will be released:

- Sample set with manual annotations: 50 documents for Scenario 1, with 100 scenario-related entity types annotated
- Source collection for Evaluation: 300K documents
- Core subset for Evaluation: 500 documents for Scenario 2, with 100 scenario-related entity types manually annotated as ground truth

Each data set includes 40% news, 40% discussion forum and 20% tweets. All documents will be in English in 2018, and additional languages will be added in subsequent years.

- **KB**

We will be using the baseKB as our target KB, as in previous years.

5 Evaluation Conditions and Schedule

Evaluation Window 1 (October 8-October 10): Each system should generate output for all of the 7,297 entity types;
Evaluation Window 2 (October 11-October 12): Systems know the target scenario-related 100 entity

types, and produce improved output;
October 15-October 19: Human analysts analyze each system output and produce corrected results for 20 unique mentions for which system produced lowest confidence values. The corrected output will be provided in the same format described in section 2.
Evaluation Window 2 (October 22-October 24): Systems incorporate human feedback and produce improved output.

6 Scoring Metric and System API Deliveries

We will apply various scoring metrics from KBP2017 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/kbp2017.pdf The scorers are posted at http://nlp.cs.rpi.edu/kbp/2017/scoring.html. If a system generates a more coarse-grained type than ground truth, it will receive partial credit based on the distance from the ground truth type in the ontology tree. We will announce detailed modification of this typing measurement metric and updated scorer soon.

This year we require each system to submit docker containers to the docker hub. NIST will collect a docker copy of each system, and run it to generate results on 500 core evaluation documents. In addition to the quality metrics, we will also measure system performance in terms of speed, memory, and storage.

7 Resource Limit

No online web search is allowed for the official run.

8 Schedule

- May 8: Release task spec
- July 1: release sample set annotations
- July 15: Registration deadline
- **October 8-24: Evaluation Window**
- October 27: Release scores to individual participants
- October 31: Participants short system description and oral presentation proposals due at NIST (for coordinators' overview paper)
- November 1: Notification of acceptance of oral presentation proposals
- November 10: Coordinator's overview paper & Participants' full workshop papers due at NIST
- November 13-14: TAC 2018 Workshop
- February 15, 2019: System description paper camera ready

9 Mailing List and Website

The KBP 2018 Entity Discovery and Linking website is http://nlp.cs.rpi.edu/kbp/2018. 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/2018/mailing.html.

10 Organizing Committee

Heng Ji (**Coordinator**, Rensselaer Polytechnic Institute, jih@rpi.edu)
Avirup Sil (**Coordinator**, IBM Research, avi@us.ibm.com)
Hoa Trang Dang (U.S. National Institute of Standards and Technology, hoa.dang@nist.gov)
Alan J. Goldschen (U.S. Department of Defense, ajgolds@tycho.ncsc.mil)
Jason Duncan (MITRE, jduncan@mitre.org)
Jeremy Getman (Linguistic Data Consortium, jgetman@ldc.upenn.edu)
Joel Nothman (University of Sydney, joel@it.usyd.edu.au)
Boyan Onyshkevych (U.S. Department of Defense, boyan.onyshkevych@darpa.mil)
Ian Soboroff (U.S. National Institute of Standards and Technology, ian.soboroff@nist.gov)
Stephanie Strassel (Linguistic Data Consortium, strassel@ldc.upenn.edu)

11 Appendix: More Examples of Extended Entity Types

- Hormone
  - ***Briain-derived neuotrophic factor*** ("***BDNF***"), another important gene in neural plasticity, has also been shown to have reduced methylation and increased transcription in animals that have undergone learning.
- Infectious Disease
  - Notable exceptions include the ***Large Pine Weevil*** ("***Hylobius abietis***"), which can kill young conifers.
- Dumpling
  - ***Shengjian mantou*** is a type of small , pan - fried " baozi " (steamed buns) which is a specialty of Shanghai.
- Fairy
  - The background story of the game starts somewhere in the desert where Anwar , a pure hearted young man finds a rusty oil lamp from what he releases a very powerful and evil ***djinn*** the Nadir.
- Lawsuit
  - The landmark ***Brown v. Board of Education*** decision paved they way for PARC v. Commonwealth of Pennsylvania and Mills vs. Board of Education of District of Columbia, which challenged the segregation of students with special needs.
- Mental Disorder
  - Many of these veterans suffer from post ***traumatic stress disorder***, an anxiety disorder that often occurs after extreme emotional trauma involving threat or injury.
- Military Academy
  - The year after, the prince went back to France,[2] where he eventually entered the prestigious academy of ***École spéciale militaire de Saint-Cyr-Coëtquidan***.
- Military Uniform
  - The following below depicted gallery of mounting loops are practically in use in conjunction with the ***5- or 3 color flectarn*** fighting suit.
- Fundraiser
  - The U.S. Fund administers the long-running ***Trick-or-Treat for UNICEF*** compaign which began as a local fundraising event in Pennsylvania in 1950 and has since raised more than US $170 million to support UNICEF's work.
- Investigator
  - Samuel Hume was born in San Francisco, California in 1885, the son of ***James B. Hume***, a famous Wells Fargo detective.
- Lobbyist

  - Represented by ***Lanny Davis***, the CES lobbied for changes to the “gainful employment rule”.
- Medical Scientist
  - Pillemer was born on October 15, 1954, to Jean Burrell Pillemer and ***Louis Pillemer***, and early pioneer in the filed of immunology at Case Western Reserve University.
- Molecular Biologist
  - Meanwhile an overlapping class of transposable element was described under the name " polintons", derived from the key proteins polymerase and integrase, by ***Vladimir Kapitonov*** and ***Jerzy Jurka***.
- Natural Language
  - The ***Vai*** language , also called ***Vy*** or ***Gallinas*** , is a Mande language spoken by the Vai people , roughly 104,000 in Liberia , and by smaller populations , some 15,500 , in Sierra Leone.
- Naval Commander
  - His ship drifting dangerously inshore , at 14:30 Captain ***Thomas Frederick*** gave control to a sailor on board who claimed to have navigated the region and knew a safe anchorage.
- Naval Gun
  - She carried one***15 cm SK L/45 gun***, four ***10.5 cm SK L/45 guns***, four ***SK L/45 gun***, four ***8.8 cm SK L/35 guns***, five ***8.8 cm SK L/30 guns***, and one ***8.8 cm SK L/30 gun*** in a U- boat mounting.
- Poisoner
  - It began to be used for murderers who used poisons after the Bishop of Rochester 's cook , ***Richard Rice*** , gave a number of people poisoned porridge , resulting in two deaths in February 1532.
- President
  - Founder 's Day is national public holiday observed in Ghana to mark the birthday of Ghana 's first president , Dr. ***Kwame Nkrumah*** the key founding father of Ghana.
- Queen
  - He later became the King of Spain and married twice to ***Marie Louise of Savoy*** and then ***Elisabeth Farnese***.
- Religion
  - The ***Mu'tazila*** tradition of tafsir has received little attention in modern scholarship , owing to several reasons.
- Salad
  - ***Texas caviar*** is a salad of black - eyed peas lightly pickled in a vinaigrette - style dressing , often eaten as a dip accompaniment to tortilla chips.
- Seafood
  - ***Lauriea siagiani***, is a species of ***squat lobster*** in the family Galatheidae, genus " ***Lauriea***".
- Sign Language
  - Following this , he has been at many festivals , including Ferstival Clin d’Œil throughout Europe as an actor , performer in various sign languages like ***DGS*** , ***BSL*** , ***LIS*** and ***LSF***.
- Appetizer
  - This invention of a faux Polynesian experience is heavily influenced by Don the Beachcomber , who is credited for the creation of the ***"pūpū" platter*** and the drink named the " Zombie " for his Hollywood restaurant.
- Bomber

  - The carburetor intake was much larger , a long duct like that on the Nakajima B6N Tenzan was added , and a large spinner — like that on the ***Yokosuka D4Y Suisei*** with the Kinsei 62—was mounted.
- Vector
  - 一般的， 令D是作用于黎曼流形M上的 ***向量丛*** V的一阶微分算子。 (In general, let D be the first-order differential operator of ***the vector bundle*** V acting on the Riemannian manifold M.)
  - 柯西 - 施瓦茨不等式叙述， 对于一个 ***内积空间*** 所有向量 " x " 和 " y " (Cauchy - Schwarz inequality description, for ***an inner product space*** of all vectors "x" and "y")
- Footbridge
  - 而较高的一座哥特式塔楼于1357年与 ***查理大桥*** 一起由彼得帕尔莱勒兴建， 直到1464年才完成。 (The taller Gothic tower was built in 1357 by Peter Parleler with the ***Charles Bridge*** until 1464.)
  - 而中国最着名的铁索吊桥是四川省甘孜的 ***泸定桥*** 。 (The most famous iron suspension bridge in China is ***Luding Bridge*** in Garze, Sichuan Province.)