TAC KBP2020 Recognizing Ultra Fine-grained EntitieS Task (RUFES)

Version 1.0 of September 1, 2020

1 Motivations and Goals

When experts query a knowledge base, their questions often involve fine-grained knowledge elements, such as "Which epidemiologists have attended the meeting?" and "Which **animo acids** in glycoprotein are most related to Glycan?". In order to answer these questions in scenarios such as disaster relief and technical support, we need to significantly extend entity extraction capabilities to a wider variety of fine-grained entity types (e.g., technical terms, lawsuits, disease, crisis, vehicles, food, biomedical entities) and to document-level knowledge aggregation. There have been many great efforts in the community to extract fine-grained entities. However, the existing benchmarks have the following limitations:

- The amount of gold-standard annotation is very small while most annotations are 'silver-standard', derived from automatic mapping from knowledge bases onto unstructured texts.
- Most data sets are Wikipedia articles instead of other documents in the real-world that may contain unknown entities.
- Given a certain context (e.g., a paragraph), an entity mention often has multiple fine-grained entity types, but existing efforts only assign one type per entity mention.
- The existing ontologies are not designed based on the distributions of entities in data from a wide range of topics.
- Existing methods have focused on name mentions only.
- There are very few shared open-source systems.

To address these limitations, we are creating a new shared task about Recognizing Ultra Fine-grained EntitieS (RUFES), extending fine-grained entity extraction to include the following innovations:

- A new task definition and a new scoring metric that requires a system to extract multiple types for each entity given an entire document as context.
- Release a new fully annotated Washington Post news corpus that consists of approximately 300 documents (50 sample, 250 test).
- Define and manually curate a new fine-grained entity ontology that consists of approximately 200 fine-grained entity types that are representative of news data.
- Extend the types of mentions to name, nominal, and pronominal mentions.

- The gold-standard annotation will include within-document coreference of entity mentions.
- A comprehensive code repository that includes open-source codes and docker containers from participants.

2 Task Overview

Given an input document, a system is required to automatically identify an entity as a cluster of name, nominal, and/or pronominal mentions, and classify the entity into one or more of the types defined in the ontology (section 3). This year we only focus on document-level entity discovery and only on English source documents. In future years we expect to add source documents in non-English languages and extend the task to corpus-level coreference. The output contains one line for each entity mention where each line has the following tab-delimited fields:

Filed 1: system run ID
Field 2: mention ID: unique for each entity name, nominal, or pronominal mention.
Field 3: mention string: the full name of a named mention, or the single head string of a nominal or pronominal mention.
Field 4: mention justification: an ID for a document in the source corpus from which the mention was extracted, the starting UTF-8 character offset of the mention, and the ending UTF-8 character offset of the mention; in the format: <document ID>:<mention start offset>-<mention end offset>
Field 5: entity ID: unique for each entity in the document.
Field 6: entity types: a set of type indicators for the mention, multiple types are separated by ";" delimiter.
Field 7: mention type: "NAM" (for name mentions), "NOM" (for nominal mentions), or "PRO" (for pronominal mentions).
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.

Offset Calculation and Formatting: KBP 2020 RUFES source documents are released in both LTF xml format and RSD ("raw source data") text format. The LTF xml file contains automatic word tokenization and paragraph segmentation, but the "RSD" text file is what the human annotator sees when annotating the document (one RSD text file per document). All annotations and offsets are with respect to the RSD text files, rather than the LTF; however, mention extents in the annotations always align with word boundaries in the LTF files.

To calculate offsets for entity mentions, the entire rsd.txt file for a given document is read into memory and represented as a UTF-8 character array, with the first character in the document

being at offset 0.[1] The start offset of a mention must be the index of the first character in the mention string, and end offset must be the index of the last character of the mention string (therefore, the length of the 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:
https://tac.nist.gov/2020/KBP/RUFES/tools.html
All documents will be in English in 2020 evaluation, and additional languages will be added in subsequent years.

3 Entity Ontology

NIST is developing a new ontology with approximately 200 fine-grained entity types. While it inherits some familiar types from previous entity-related evaluations, the focus is on new fine-grained types that are relevant and salient to the topics of selected articles. They will follow the same three level x.y.z hierarchy as in the TAC-KBP 2019 EDL track, but some modifications have been made to a few inherited entity types.

Entity types are being developed in conjunction with article selection using a tool developed for TREC News Track topic research (though one should not assume that topics in this task are related to topics in the TREC News Track). NIST staff use keywords to query the corpus and read one or more articles carefully from the list of returned articles. For each mention deemed salient to the article, if the entity it refers to does not fall into any fine-grained entity type of the existing ontology, NIST staff research YAGO and Wikidata to create a new entity type and add it to the ontology set. Additional "sister" entity types may be added if they are potentially relevant to the topic. System output from two UIUC taggers run on the entire corpus, one using the YAGO ontology and the other the AIDA ontology, is also used to assist in ontology development.

Each entity type will have one sentence definition along with some examples. Mappings to YAGO will be included if available. Note that due to different granularities, some mappings are one to many.

4 Data Sets

4.1 Source Data

The KBP 2020 RUFES development source documents and evaluation source documents will be drawn from a collection of Washington Post news articles from January 2012 to December 2019.

[1] Note that in documents with non-ASCII content – i.e. multi-byte (wide) characters in UTF-8 -- the offsets refer to characters, not bytes.

The development source corpus and the evaluation source corpus will each comprise approximately 100,000 unannotated articles from the Washington Post. Participants must refrain from training/developing their systems on any Washington Post articles from 2012 or later, except for those articles released by NIST.

NIST will select around 300 articles (about 150,000 words) from the evaluation source corpus for annotation. We will split them into 50 documents for development and 250 documents for evaluation. Selected articles may cover a variety of topics, from domestic politics to international relations, from regulation of consumer products to financial market trends, from advance in medicine to space programs, and so on. Lengthy articles will be truncated at a natural paragraph boundary to about 500 words. We will also organize a community effort to annotate more data following the ontology to assist system development.

4.2 Annotation

Each article will be annotated by one annotator. A second pass will be performed on each annotated article by a different annotator for quality assurance.

Named, nominal, and pronominal mentions will be annotated and each mention will be labeled as either NAM, NOM, or PRO. For a named mention, the mention span includes the full extent of the mention excluding any pre- or post-modifiers because a modifier is not an integral part of the name. Determiners are also excluded even if they are capitalized in the middle of a sentence.

For nominal and pronominal mentions, only the head word is tagged. This also applies to space delimited compounds where only the last token is tagged. This simple approach is to help ensure annotation consistency.

No embedded mention within the span of a named mention will be tagged. For nominal mentions, mentions appearing as (part of) a modifier will be tagged since the nominal mention span only extends to the head and no embedded mention spans will be produced.

For each article, the annotator will tag each taggable entity mention and assign the most fine-grained entity types to the mention that can be determined from the local context, which is defined as the paragraph in which the mention appears. The annotator can back off to more coarse-grained types if needed. There is no need to read beyond the paragraph for typing.

Entity coreference within a document is done by associating each mention to a “canonical” form of the entity it refers to. The annotator may use the first named mention of an entity as the canonical form or type in a canonical form of their choice. Each canonical form must be unique and unambiguous within the article.

4.3 Additional training and development data

There are many related data annotations from previous years in TAC-KBP that can be used to develop and improve systems. Participants may request any of the LDC packages when registering for the TAC KBP 2020 RUFES task. The full list of corpora is listed in Appendix A. Here we highlight some corpora:

- LDC2014T16 TAC KBP Reference Knowledge Base
- LDC2015E42 TAC KBP Knowledge Base II – BaseKB
- LDC2018T16 TAC KBP English Entity Linking - Comprehensive Training and Evaluation Data 2009-2013
- LDC2019T02 TAC KBP Entity Discovery and Linking - Comprehensive Training and Evaluation Data 2014-2015
- LDC2019T19 TAC KBP Entity Discovery and Linking - Comprehensive Evaluation Data 2016-2017
- LDC2019E78 TAC KBP 2019 EDL Track Fine-Grained Name Tagging Evaluation Annotations
- LDC2019E79: TAC KBP 2019 EDL Track Fine-Grained Name Tagging Evaluation Source Corpus: 300K source documents for the EDL 2019 Evaluation. Participants output entity types for all 300K documents, but only a core subset of the documents were evaluated (in LDC2019E78).
- **UIUC silver-standard annotation from Wikipedia and YAGO**
    - UIUC created silver-standard annotation derived from Wikipedia markups [Pan et al., ACL2017] for 16K+ YAGO entity types (571GB): https://drive.google.com/file/d/1Bh7VZvgo9N_BVN9aRvAqoKhcavep4NNW/view
- **UIUC human annotation for DARPA AIDA seedling data in 2019:**
    - http://blender.cs.illinois.edu/kbp/2019/KBP2019-UIUCAnnotation.tar.gz
- **Other Related Resources:**
    - FIGER: https://github.com/xiaoling/figer (derived from Wikipedia)
    - Ultra-fine grained + augmented OntoNotes: https://github.com/uwnlp/open_type
    - KNET: https://github.com/thunlp/KNET (derived from Wikipedia

5 Scoring Metrics

We will report various coreference scoring metrics from the KBP2017 EDL task and also report a new entity-level type metric to evaluate fine-grained entity typing. The detailed description of the coreference metrics is in section 2.2 in the overview paper: http://nlp.cs.rpi.edu/paper/kbp2017.pdf The scorers are posted at http://blender.cs.illinois.edu/kbp/2019/scoring.html

We design the following metric for entity typing:

Each system entity ID (Field 5) has a set of system types (Field 6) associated with it. The set of types associated with an entity will be the union of the set of types asserted for any mention

(Field 2) of that entity, plus more coarse-grained types (up to the top-level type for entities). For example, if the system entity has mentions with types {PER.MilitaryPersonnel, PER.Politician.Mayor}, then the set of types used for the evaluation will be {PER, PER.MilitaryPersonnel, PER.Politician, PER.Politician.Mayor} for that entity. The set of types for each gold entity is constructed in the same way from the gold annotations.

Using the alignment of system entity IDs and gold entity IDs from mention-level CEAF, NIST will compute Type Precision, Type Recall, and type F1 on the types for each pair of aligned system and gold entities; unaligned system entities and gold entities each have F1=0. For the final score, NIST will report Macro-Averaged Type F1, which is the mean F1 over the unaligned entities and the pairs of aligned entities.

6 Human Feedback

Human feedback will be provided to systems based on a user model of how analysts might interact with the system. In the model, the user is reviewing a document and working through it from the beginning to the end. During this sequential process, if at time t the user encounters an incorrect system-generated annotation (which could be a miss, an error in mention extent, a type error, or a coreference error), the user tags the error and provides the correction. The user has some tolerance value k for the amount of error they are willing to see before they lose confidence in the system's annotations and stop reviewing the document. Let the amount of error encountered so far in document d at time t be a function Err(d,t). The user will stop reviewing the document as soon as Err(d,t) > k. Err(d,t) could be a function of the number of errors seen up to time t, how egregious the errors are, and possibly how the errors are distributed across the document. For 2020, we plan to use a simple error function:

Err(d,t) = total number of errors seen up to time t

And k will be a small number (likely k=10 as in TAC-KBP2019).

7 Bonus System Submissions through Docker Copies

This year we will request participants to submit their system worksheets (via Codalab) along with the corresponding system output. This would encourage reproducibility of empirical numbers and promote future research in this direction.

We will have two separate methods for submissions:

- **Traditional submission**: participants submit results on source collection to NIST submission website
- **Bonus docker submission**: Besides the regular submission methods by uploading results to NIST website, this year we highly encourage participants to join a separate bonus track by submitting their final systems to NIST as docker containers. We hope this will help us build a

common code repository to share with the KBP community and facilitate improving the replicability of research results. We will follow the Codalab worksheets format to create a repo for the system submissions. For info: https://github.com/codalab/codalab-worksheets/wiki. Then each participant is asked to submit the worksheets to NIST and we will list them in our Workshop webpage for others to download.

Alternate route 1 (Worksheet):
Participants are asked to submit their worksheets directly to NIST. NIST will run the worksheet of each system, and run it to generate results on 250 core evaluation documents. In addition to the quality metrics, we will also measure system performance in terms of speed, memory, and storage.

Alternate route 2 (Docker):
Participants are asked to submit their docker containers to the docker hub. NIST will collect a docker copy of each system, and run it to generate results on 250 core evaluation documents. In addition to the quality metrics, we will also measure system performance in terms of speed, memory, and storage.

UIUC team will develop a baseline system [Lin et al., EMNLP2019] using the available resources and basic embedding features, and make the source code publicly available for participants. All participants should aim to outperform this baseline by novel approaches. After the evaluation when we select teams to give oral presentation, we will also evaluate the novelty and soundness of a team's method, instead of just score ranking.

8 Schedule

- September 1: Release task spec
- September 15: Release ontology
- September 15: Release development source corpus (100,000 unannotated Washington Post documents for embedding learning and development)
- November 1: TAC 2020 Track Registration deadline
- November 15: Release sample annotations on 50 documents
- December 1: Release evaluation source corpus (100,000 Washington Post documents)
- **December 7-December 20: Evaluation Windows**
  - **December 7-December 11 (Evaluation Window 1):** Systems generate output for the entity types that are in the KBP 2020 RUFES ontology.
  - **December 13**: Humans provide feedback for system output. The format of feedback is the same format as RUFES system output.
  - **December 13-December 20 (Evaluation Window 2):** Systems incorporate human feedback and produce improved output. For 2020 we will ask annotators to simulate

human feedback. In future years we expect to provide human feedback from analysts instead of manual annotations.

- December 23: Release scores to individual participants
- January 5: Participants short system description and oral presentation proposals due at NIST (for coordinators' overview paper)
- January 10: Notification of acceptance of oral presentation proposals
- January 15, 2021: Coordinator's overview paper & Participants' full workshop papers due at NIST
- January 25-26, 2021: Thirteenth TAC Workshop (online)
- March 1, 2021: System description paper camera ready (final proceedings version)

9 Mailing List and Website

The KBP 2020 RUFES website is https://tac.nist.gov/2020/KBP/RUFES/ . Please post any questions and comments to the mailing list tac-kbp@list.nist.gov. Information about subscribing to the list is available at: https://tac.nist.gov/2020/KBP/RUFES/

10 Organizing Committee

Heng Ji (Coordinator, University of Illinois at Urbana-Champaign, hengji@illinois.edu)

Avi Sil (Coordinator, IBM Research AI, avi@us.ibm.com)

Hoa Trang Dang (U.S. National Institute of Standards and Technology, hoa.dang@nist.gov)

Shudong Huang (U.S. National Institute of Standards and Technology, shudong.huang@nist.gov)

Joel Nothman (University of Sydney, joel@it.usyd.edu.au)

Ian Soboroff (U.S. National Institute of Standards and Technology, ian.soboroff@nist.gov)

11 Scientific Board

Hanna Hajishirzi (University of Washington)

Ed Hovy (U.S. Department of Defense)

Yunyao Li (IBM Research AI)

Andrew McCallum (University of Massachusetts Amherst)

Paul McNamee (Johns Hopkins University)

Graham Neubig (Carnegie Mellon University)

Boyan Onyshkevych (U.S. Department of Defense)

Marius Pasca (Google)

Siddharth Patwardhan (Apple)

Dan Roth (University of Pennsylvania)

Xiang Ren (University of Southern California)

Satoshi Sekine (RIKEN Center for Advanced Intelligence)

Sameer Singh (University of California Irvine)

Appendix A: Full List of Related Corpora Available to Participants from LDC

LDC2013E79 DEFT - Phase 1 Textual Entailment Annotation R1
LDC2014E07 DEFT - Phase 1 Textual Entailment Annotation R2
LDC2014E38 DEFT - Phase 1 Inference Annotation Pilot
LDC2014T16 TAC KBP Reference Knowledge Base
LDC2015E29 DEFT Rich ERE English Training Annotation
LDC2015E42 TAC KBP Knowledge Base II - BaseKB
LDC2015E47 TAC KBP English Sentiment Slot Filling - Comprehensive Training and Evaluation Data 2013-2014
LDC2015E49 TAC KBP English Surprise Slot Filling – Comprehensive Training and Evaluation Data 2010
LDC2015E68 DEFT Rich ERE English Training Annotation R2
LDC2016E114 TAC KBP 2016 Belief and Sentiment Evaluation Gold Standard Annotation
LDC2016E27 DEFT English Belief and Sentiment Annotation
LDC2016E31 DEFT Rich ERE English Training Annotation R3
LDC2016E73 TAC KBP 2016 Eval Core Set Rich ERE Annotation with Augmented Event Argument
LDC2016E94 LORELEI Russian Representative Language Pack Monolingual Text
LDC2017E02 TAC KBP Event Nugget Detection and Coreference - Comprehensive Training and Evaluation Data 2014-2016
LDC2017E05 TAC KBP Event Argument Comprehensive Training and Evaluation Data 2014-2016
LDC2017E53 TAC KBP 2017 Eval Core Set Rich ERE Annotation
LDC2017E54 TAC KBP 2017 Eval Core Set Event Nugget Annotation

LDC2017E55 TAC KBP 2017 Eval Core Set Rich ERE Annotation with Augmented EventArguments
LDC2017E79 TAC KBP 2017 Event Sequencing Eval Source Data
LDC2017E80 TAC KBP 2017 Belief and Sentiment Evaluation Gold Standard Annotation
LDC2017E83 TAC KBP 2017 Event Sequencing Eval After Link Parent Child Annotation
LDC2018E01 AIDA Scenario 1 - Seedling Corpus V2.0
LDC2018E19 LoReHLT Russian Representative Language Pack Translation, Annotation, Grammar, Lexicon and Tools
LDC2018E45 AIDA Scenario 1 - Seedling Annotation V6.0
LDC2018E52 AIDA Scenario 1 - Seedling Corpus Part 2 V2
LDC2018E53 AIDA Scenario 1 - Seedling Background Corpus Non-Eval
LDC2018E63 AIDA Scenario 1 - Seedling Corpus Training Data Video Segmentation V2.0
LDC2018E64 AIDA Scenario 1 - Seedling Background Corpus Non-Eval Video Segmentation V1.0
LDC2018E76 AIDA Month 9 Pilot Eval Annotation - Unsequestered V1.0
LDC2018T03 TAC KBP Comprehensive English Source Corpora 2009-2014
LDC2018T16 TAC KBP English Entity Linking - Comprehensive Training and Evaluation Data 2009-2013
LDC2018T22 TAC KBP English Regular Slot Filling - Comprehensive Training and Evaluation Data 2009-2014
LDC2019E04 AIDA Phase 1 Evaluation Practice Topic Source Data V2.0
LDC2019E07 AIDA Phase 1 Evaluation Practice Topic Annotations
LDC2019E44 AIDA Phase 1 Practice Topics Reference Knowledge Base
LDC2019E77 AIDA Phase 1 Evaluation Topic Annotations Unsequestered V2.0
LDC2019E78 TAC KBP 2019 EDL Track Fine-Grained Name Tagging Evaluation Annotations
LDC2019E79 TAC KBP 2019 EDL Track Fine-Grained Name Tagging Evaluation Source Corpus
LDC2019T02 TAC KBP Entity Discovery and Linking - Comprehensive Training and Evaluation Data 2014-2015
LDC2019T09 DEFT Spanish Committed Belief Annotation
LDC2019T12 TAC KBP Evaluation Source Corpora 2016-2017
LDC2019T16 DEFT English Committed Belief Annotation
LDC2019T17 TAC KBP Cold Start - Comprehensive Evaluation Data 2012-2017
LDC2019T17 TAC KBP Cold Start - Comprehensive Evaluation Data 2012-2017
LDC2019T17 TAC KBP Cold Start - Comprehensive Evaluation Data 2012-2017
LDC2019T19 TAC KBP Entity Discovery and Linking - Comprehensive Evaluation Data 2016-2017
LDC2020T08 TAC KBP English Temporal Slot Filling - Comprehensive Training and Evaluation Data 2011 and 2013