TAC KBP 2022 Recognizing Ultra Fine-grained EntitieS Task (RUFES)

Version 1.0 of October 1, 2022

1 Motivations and Goals

When experts query a knowledge base, their questions often involve entities with very fine-grained types, such as "Which **epidemiologists** have attended the meeting?" and "Which **amino acids** in glycoprotein are most related to Glycan?". In order to answer specific 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., different types of vehicles, various diseases and biomedical entities, etc.) and to document-level knowledge aggregation. There have been many excellent efforts in the community to extract fine-grained entities. However, the 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 sometimes has multiple fine-grained entity types, but most 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.
- Most methods have focused on proper name mentions only.
- There are very few shared open-source systems.

To address these limitations, a shared task about Recognizing Ultra Fine-grained EntitieS (RUFES) was created in 2020, extending fine-grained entity extraction to include the following innovations:

- Defined a new task and scoring metric that require a system to extract multiple types for each entity given an entire document as context.
- Defined and manually curated a new fine-grained entity ontology that consists of approximately 200 fine-grained entity types that are representative of news data.
- Extended the set of taggable entity mentions to name, nominal, and pronominal mentions.
- Included within-document coreference of entity mentions in the gold-standard annotation.
- Released a new fully annotated Washington Post news corpus that consists of over 150 documents.

TAC KBP 2022 extends the original RUFES task in the following ways:

- Manually extend the entity ontology to include approximately 300 fine-grained entity types.
- Extend the set of taggable entity mentions to include mentions that make generic references as well as specific references to entities.
- Fully annotate a Washington Post news corpus that consists of 350 documents that have been annotated according to the new ontology and annotation guidelines.
- Introduce a leaderboard for participants to improve their systems and track evaluation results.

2 Task Overview

RUFES requires systems to perform three subtasks: Mention Extraction, Entity Coreference Resolution, and Fine-grained Entity Typing. Given an input document, a RUFES system is required to automatically identify an entity as a cluster of mentions (comprising name, nominal, and/or pronominal mentions), and classify the entity into one or more of the types defined in the ontology (section 3). In the 2022 evaluation, all documents are in English, and we focus only on document-level clustering of entity mentions. In future years we expect to add source documents in non-English languages and extend the task to corpus-level coreference.

The RUFES system must process all documents in the evaluation source corpus and output a single file for the entire evaluation source corpus. 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 name, nominal, or pronominal entity mention.
Field 3: mention string: the full name of a named mention, or the 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; mention justification has 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, selected from the entity ontology; multiple types are separated by the ";" 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 2022 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 tool to validate RUFES submission file format is available at:
https://tac.nist.gov/2022/KBP/RUFES/tools.html

3 Entity Ontology

NIST has developed an ontology with approximately 300 fine-grained entity types for RUFES. While the ontology inherits some familiar types from previous entity-related evaluations, the focus is on fine-grained types that are relevant and salient to the topics of selected articles. The ontology follows 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.

RUFES entity types are developed in conjunction with article selection. 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 Wikidata and YAGO to create a new entity type and add it to the ontology. Additional "sister" entity types may be added if they are potentially relevant to the topic. System output from two automatic taggers is used to assist in ontology development.

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

[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.

4 Data Sets

4.1 Development and Evaluation Data

The KBP 2022 RUFES development source documents and evaluation source documents are drawn from a collection of Washington Post news articles from January 2012 to December 2020. Lengthy articles are truncated at a natural paragraph boundary so that the average document length is 500 words.

The following KBP 2022 RUFES data are available for download from the data page of the RUFES website:

- TAC KBP2022 RUFES Development Data Sets: The development data comprises approximately 200,000 unannotated documents (for embedding learning and development) and annotations for 100 of these documents. The development source corpus is a re-release of the 2020 development set, plus 100,000 new development articles for 2022. The sample annotations consist of 50 new fully annotated documents and 50 sample documents that were released in 2020 but that have been reannotated using the ontology and annotation guidelines for 2022.
- TAC KBP2022 RUFES Evaluation Data: The evaluation source corpus comprises 10,000 articles.

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 has selected and annotated 250 documents from the evaluation source corpus. Selected articles 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. NIST will use the annotations on 25 of those documents to provide feedback to participants, while the remaining 225 documents will be used as the gold standard for evaluating systems participating in the TAC KBP 2022 RUFES task.

4.2 Annotation

Each article is annotated by one annotator, and a second pass is performed on each annotated article by a different annotator for quality assurance.

Annotators annotate named, nominal, and pronominal entity mentions and label each mention 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 head is tagged. However, for hyphenated compounds ("editor – in – chief", "hunter – gatherer"), the entire compound is tagged as a single mention head (in part because headless compounds such as "hunter – gatherer" tend to be written with hyphens).[2] This simple approach is to help ensure annotation consistency.

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

For each article, the annotator tags each taggable entity mention and assigns the most fine-grained entity types to the mention that can be determined from the local context, which is defined to be 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. While TAC KBP 2020 RUFES data can be downloaded directly from NIST, earlier TAC KBP data must be requested from the LDC. Here we highlight some corpora that participants may request from the LDC when registering for the TAC KBP 2022 RUFES task, as well as other related resources:

- Past TAC KBP datasets from the LDC:
    - 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

[2] The treatment of hyphenated compounds in 2022 differs from 2020.

output entity types for all 300K documents, but only a core subset of the documents were evaluated (in LDC2019E78).

- Past TAC KBP data from NIST:
    - TAC KBP 2020 RUFES Sample Annotations V1.1
    - TAC KBP 2020 RUFES Evaluation Data
    - TAC KBP 2020 RUFES Evaluation Annotations
- 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

NIST will report various diagnostic metrics for detection of mentions and their mention-level types (fine-grained types, as well as top-level types), and for clustering of coreferential mentions. For the final score, we will report the RUFES ClusterTypesMetricV1, an entity-level type metric (first defined for RUFES 2020) to evaluate performance on the overall task of coreference and fine-grained entity typing. All metrics require a system mention span to be identical with a gold mention span in order for the mentions to be considered a match.

Evaluation metrics for TAC KBP 2022 RUFES task:

- For mention detection, coarse-grained type classification, and coreference (https://neleval.readthedocs.io/en/latest/basic_measures.html):
    - Strong_mention_match: mention detection
    - Strong_typed_mention_match: mention detection and type classification (only top level type must match)
    - Mention_ceaf: clustering
    - Typed_mention_ceaf: clustering and type classification (only top level type must match)
    - Entity_ceaf: clustering
- For the combined task of mention detection and fine-grained type classification, with and without coreference (https://github.com/shahraj81/rufes):
    - MentionTypesMetricV1: mention-level fine-grained types
    - ClusterTypesMetricV1: clustering and entity-level fine-grained types.

The RUFES ClusterTypesMetricV1 is computed as follows:

1. For each document, construct the set of types for each system entity in the document: 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 is defined to 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.

2. For each document, construct the set of types for each gold entity in the document: The set of types for each gold entity is constructed from the gold annotations in the same way as the set of types for each system entity.
3. For each document, using the alignment of system entity IDs and gold entity IDs from mention 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.
4. For the final ClusterTypesMetricV1 score, NIST reports Macro-Averaged Type F1, which is the mean F1 over the unaligned entities and the pairs of aligned entities.

The RUFES MentionTypesMetricV1 is computed in the same way as the ClusterTypesMetricV1, except that each mention is treated as a cluster with just a single mention.

Note that ClusterTypesMetricV1, measuring performance on entity-level fine-grained types, does not require the system to return all mentions of an entity. To get a perfect score for this metric, the system must find all entities that have a type in the ontology, but does not need to find all mentions of each entity; it is sufficient for the system to find a subset of mentions for each entity, such that the mentions cover the range of types that are attested for the entity in the document.

Scorers are posted at https://tac.nist.gov/2022/KBP/RUFES/tools.html

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 missed or spurious mention, 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 2022, we plan to use a simple error function:

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

Human feedback will be given for 25 documents and will be in the same format as the gold standard annotation, except that no annotations will be provided after the first k=10 errors have been encountered in the document. The feedback documents are selected to target some entity types that have few or no annotated examples in previously released RUFES annotations.

7 Submissions

We are introducing a leaderboard for the TAC KBP 2022 RUFES evaluation to allow participants to submit multiple runs and track evaluation results.

For each run, the participant will run a particular configuration of their RUFES system on all the documents in the TAC KBP2022 RUFES Evaluation Data and upload a single system output file to the leaderboard. If there are any format errors encountered in the submission file, the submission will be rejected in its entirety, and no attempt will be made to score it. Therefore, before uploading their RUFES system output file to the leaderboard, the participant must verify that the file has valid format, using the RUFES submission validator at: https://tac.nist.gov/2022/KBP/RUFES/tools.html

Each participant team is allowed to submit up to **10 runs per month** to the leaderboard, to **2 runs per day**.

In addition to submitting system output to the leaderboard, participants are highly encouraged to submit 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. NIST will also run the dockers to measure system performance in terms of speed, memory, and storage.

8 Schedule

The evaluation schedule is posted on the TAC KBP 2022 RUFES track web site: https://tac.nist.gov/2022/KBP/RUFES/schedule.html

9 Mailing List and Website

Track coordinators will send all announcements about the RUFES task and evaluation to the mailing list tac-kbp@list.nist.gov. Participants in the RUFES task must subscribe themselves to the mailing list. Information about subscribing to the list is available on the KBP 2022 RUFES website: https://tac.nist.gov/2022/KBP/RUFES/ .

10 Organizing Committee

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

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

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

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