NIST LoReHLT 2018 Evaluation Plan

Last Updated: June 7, 2018

Revision History

June 7, 2018 (v1.0.4)

- Fixed typos in SF section
- Fixed typos and formatting in schedule

May 31, 2018 (v1.0.3)

- Situation Frame section corrections
    - Required fields modified
    - Field names corrected to match the scorer
    - Clarified need/issue frame requirements and status reporting difference (need: current/past/future; VS. issue: current/non-current)
    - Example json output corrected to match the schema
    - Json schema version specified
    - Corrected conversion protocol from scope/severity to urgency

April 20, 2018 (v1.0.2)

- Situation Frame section rewrite
    - nDCG as primary metric
    - Diagnostic metrics include PR curves
    - Relationship between “Urgency” and the new LDC annotation fields of “scope” and “severity” explained

April 4, 2018 (v1.0.1)

- Fixed typo in main schedule
- Incorporated WERB comments (minor fixes)
- Clarified SF/MT annotation description

March 2, 2018 (v1.0.0)

- Updated the registration process
- Updated the definition of ensemble
- Fixed error in description of Constrained Training condition
- Updated MT section to specify that the English portion of set E may not be used in any way for MT system development
- Updated MT performance measurement section
- Updated EDL section with input from DARPA: 1) specify that each IL will have its own KB; and 2) NIST report scores with and without English nominals

- Removed Somali and Yoruba from OTAL EDL and Chinese and Arabic from OTAL MT. OTAL task will only use IL5 and IL6
- Added PI meeting dates to schedule
- Edited MT scoring section to clarify justification segments and add (potential) human assessment
- Changed main LoReHLT18 evaluation period to Jul 2 - 11
- Changed OTAL task evaluation period to Sep 17 - 21
- Changed registration period to Mar 1 - May 31
- Minor format & cosmetic fixes

January 17, 2018 (v0.0.0)

- Initial release

1 Introduction

The 2018 Low Resource Human Languages Technologies (LoReHLT) evaluation is the third evaluation in the National Institute of Standards and Technology (NIST) LoReHLT evaluation series that began in 2016. The series was designed in collaboration with the Defense Advanced Research Projects Agency (DARPA) Low Resource Languages for Emergent Incidents (LORELEI) Program to develop human language technology (HLT) that can support rapid and effective response to emerging incidents where the language resources are very limited. As such, LORELEI aims to develop capabilities that can extract knowledge from foreign language sources quickly. This document describes the evaluation specifications of the component evaluation conducted by NIST to assess the performance.

Similar to the 2017 evaluation, the 2018 evaluation will include two surprise languages instead of one as in 2016. However, unlike 2017 a certain amount of English data will also be included in the test set for each incident language (IL) for the situation frame (SF) and entity detection and linking (EDL) tasks. The SF task will be redefined with a focus toward recognizing collection level "situations" and with audio and text evaluated as a single task.

In 2018 the number of checkpoints will be reduced to two instead of three. Again, there will be no distinction between primary or contrastive systems, and teams can submit up to 10 submissions per checkpoint. However, no feedback score will be given on any part of the datasets.

New to LoReHLT18 is an open task with much simpler protocol (no surprise language element, over a single checkpoint, and unconstrained training[1]) with the main focus on cross language techniques.

Participation in the NIST Low Resource Human Language Technology (LoReHLT) evaluation is required for all DARPA LORELEI performers responsible for the relevant component technologies in LORELEI. The evaluation is also open to all researchers who find the evaluation tasks of interest. There is no cost to participate. However, participants are expected to attend a post-evaluation workshop to present and discuss their systems and results at their own expense. Information and updates about the component evaluation will be posted to the NIST LoReHLT website.[2]

2 Evaluation Tasks

There are four evaluation tasks. LORELEI performers are required to participate in the tasks as outlined by their Statement of Work. Open participants (non-LORELEI performers) can participate in any and all tasks.

- **Machine Translation (MT)** – given a collection of IL text documents, automatically translate them to English. For MT specific requirements, see section 15 Machine Translation (MT) Evaluation Specifications.
- **Situation Frame (SF)** – given a collection of audio and text documents in IL and English, automatically generate situation frames covered in the collection, and link those situation frames into knowledge base (KB) level situations. For SF specific requirements, see section 16 Situation Frame (SF) Evaluation Specifications.

[1] Can be anything outside of evaluation epoch.

[2] https://www.nist.gov/itl/iad/mig/lorehlt-evaluations

- **Entity Discovery and Linking (EDL)** – given a collection of text documents in IL and English, identify named mentions in both IL and English as well as nominal mentions in English, classify them into predefined entity types, and link the mentions to a KB or cluster them if they are not linkable to the KB. For EDL specific requirements, see section 17 Entity Discovery and Linking (EDL) Evaluation Specifications.
- **Open Test on Additional Languages (OTAL) –** given multiple sets of text documents in multiple previously released LoReHLT languages, run a simplified variant of the MT and EDL tasks using a common system across all languages. Since this task has a number of simplifications from the others, it is described in its own section 13 Open Test on Additional Languages.

Task	**Language**	**Input**
MT	IL9, IL10	Text
SF	IL9, IL10, English	Audio and Text
EDL	IL9, IL10, English	Text
OTAL	Oromo (MT/EDL) Tigrinya (MT/EDL)	Text

Table 1: LoReHLT18 Tasks

3 Time Machine Principle

The LoReHLT evaluation focuses on evaluating technologies that can support rapid and effective response to emerging incidents (e.g., earthquake, hurricane) in a low resource language (also referred to as incident language). As such, a portion of the evaluation data contains incident-relevant data. To make the evaluation feasible, the incident must already have happened to make data collection for system training and testing possible. To mimic that the incident has not happened yet, systems should not mine for data about the incident in any language and developers should not ask the native informant (NI) about the incident after the incident is announced as both would constitute "knowing the future". In a live situation, systems will get more information about the incident as the incident develops. This is being simulated by the additional training data teams will be given in the constrained training condition. However, this situation is harder to simulate with the native informant, so to make the evaluation easier to manage, developers are not allowed to ask the native informant about the incident[3].

Mining for all incidents from the internet (e.g., create SFs for all incidents found on the internet) would violate the time machine principle described above unless teams can categorize their incidents by date and can quickly roll back to the time before the incident, when the incident is announced[4].

[3] Please see section 7 Native Informant Resources for complete guidelines regarding the native informant usage.

[4] If teams cannot roll back, they cannot use the data in the constrained training condition. Teams will be allowed to use it in the unconstrained condition if and only if they can demonstrate performance difference due to knowledge of the future.

4 Training Conditions

For each evaluation task, there are two training conditions, constrained and unconstrained, that differentiate the amount and source of incident language-related training material without preventing or excluding multilingual resources and technologies. Prior to the incident and incident language announcement, teams can assemble multilingual resources/technologies/etc. to build their system so long as the resources are multilingual-focused in nature. Teams will be also given some resources to use; those resources are described in 4. Serendipitous inclusion of the incident language data in a multilingual system is allowed and must be documented in the system description. The use of pre-existing, mono-lingual technologies for the incident language is allowed as long as the technology is not a LoReHLT task. For instance, running the test data through GoogleTranslate™ is not permitted since MT is a LoReHLT task.

- **Constrained** – The intent of the *constrained* training condition is to test multilingual systems that are re-targeted to an incident language using a fixed set of incident language resources after the incident and the incident language are announced. The fixed set is described in section 5 Baseline Training Data, and no other incident or non-incident language materials (i.e., parallel text, speech corpora, etc.) are permitted. In addition, knowledge about the incident language gained from the Native Language Informant within the allotted time and followed the procedures outlined in section 7 Native Informant Resources is permitted. The constrained training condition is **required for each task participated in**.
- **Unconstrained** – The intent of the *unconstrained* training condition is to see performance gain when additional publicly available data are allowed (outside of what is described in section 5 Baseline Training Data). Teams can mine for additional data but should not violate the time machine principle by mining specifically for incident-related data after the incident is announced. Teams can use additional Native Informant time beyond the limits in section 7 Native Informant Resources[5]. Prior to the incident and incident language announcement, teams can assemble mono- and bilingual resources including those in the incident language. The unconstrained training condition is **optional but encouraged**.

5 Baseline Training Data

For each evaluation task, a set of non-IL data resources will be provided by the LDC for training prior to the evaluation period. To obtain this data, open participants must register to participate and sign the license agreement which can be found on the NIST LoReHLT website.

Each task (MT, SF, or EDL) has its own annotation guidelines. If you are an open participant and do not have direct access to the annotation guidelines, please contact LDC (lorelei-poc@ldc.upenn.edu) for the LoReHLT translation, situation frame, or entity discovery and linking guidelines.

[5] LORELEI performers must make prior arrangements directly with Appen if they want additional time with the native informant.

6 Evaluation Data

The LoReHLT18 will have two incident languages which will be referred as IL9 and IL10. In addition, English data will be included in the evaluation dataset (Set E). Each incident language follows the same data component and format as described below.

6.1 Component Definition & Release Plan

MT, EDL, and SF will be evaluated at all two checkpoints. The LDC will release the data in an encrypted format (see section 6.4 Data Encryption) at the Pre-IL Announcement stage, and NIST releases the appropriate decryption key(s) at the later stages listed below. Both ILs follow the same data release schedule. The stages are:

- **Pre-IL Announcement** (June 29, 2018)
  - **KB:** Encrypted knowledge base released
  - **Set 0**: Encrypted pre-incident IL training data released
  - **Set 1**: Encrypted incident/post-incident IL training data set 1 released
  - **Set S**: Encrypted incident/post-incident English Scenario Model released
  - **Set E**: Encrypted incident/post-incident IL evaluation data released
- **IL Announcement** (12:00 ET July 2)
  - Identity of IL announced (by LDC)
  - Decryption keys for **KB**, **Set 0** and **Set E** released (by NIST)
- **Evaluation Checkpoint 1** (12:15 ET July 2 - 12:00 ET July 3)
  - Train with data from **Set 0** begins
  - Submission due at the end of Evaluation Checkpoint 1
  - At the end of Evaluation Checkpoint 1, decryption keys for **Set 1** and **Set S** released
- **Evaluation Checkpoint 2** (12:15 ET July 3 - 12:00 ET July 11)
  - Train with data from **Set 1** and **Set S** begins
  - Submission due at the end of Evaluation Checkpoint 2

6.2 Data Description

The composition of the KB and datasets (**KB, Set 0, Set 1, Set S, Set E**) for each incident language are listed in Table 2. The given target data volume is approximate and depends on data availability. If the amount for a genre is short of the target, LDC will substitute another genre. “kw” refers to multiples of 1000 words.

6.3 Data Format and Structure

These datasets (**KB, Set 0, Set 1, Set S, Set E** aka the evaluation IL package) will be released by the LDC. The data format and structure are described in detail in the data specification document uploaded on the NIST LoReHLT website.

6.4 Data Encryption

The datasets listed in Table 2 will be encrypted using OpenSSL. NIST has created a package with instructions on how to encrypt and decrypt the data using some sample data. The package can be downloaded from the NIST LoReHLT website.

Set 0 – pre-incident epoch

Category I Resources[6]

- Monolingual Source Text:
  - Approx. 100 kw newswire
  - Approx. 75 kw discussion forum/blog
  - Approx. 50 kw Twitter/SMS
- Monolingual Source Speech:
  - Several hours of audio - in-domain and out-of-domain, pre-incident[7]
- Parallel Text[8]:
  - Approx. 100 kw newswire
  - Approx. 100 kw discussion forum/blog
  - Approx. 100 kw Twitter/SMS
- Parallel Dictionary (~10,000 stems/lemmas)

Category II Resources (any 5 of the following):

- parallel dictionary IL --> non-English
- monolingual IL dictionary
- monolingual IL grammar book
- parallel English --> IL grammar book
- monolingual IL primer book
- monolingual IL gazetteer
- parallel IL --> English gazetteer

Set 1 – incident/post-incident epoch

Monolingual Source Text – leftover data after **Set E** is met (maximum approx. 1.5 Mw)

Monolingual Source Speech:

- Several hours of audio - in-domain and out-of-domain, incident/post-incident

Set S – incident/post-incident epoch

English Scenario Model – up to 50 kw (text only), genre balance will vary based on availability

Set E – incident/post-incident epoch

[6] One of the category I resources (monolingual text, parallel text, or parallel dictionary) must exceed the minimum target by 500%.

[7] Set 0 and Set 1 of speech data will make up a total 14 h of audio; 60% in-domain, pre-incident and incident/post-incident data; 40% out-of-domain; 70% of formal data and 30% of informal data, +/-10% variance.

[8] The parallel text is found/harvested data and automatically aligned, not created (e.g. via professional translation agency or crowdsourcing). ~300kw comparable may be substituted for every 100kw parallel if parallel text is not available.

Source Text:
• Approx. 100 kw newswire • Approx. 50 kw discussion forum/blog • Approx. 50 kw Twitter/SMS • Approx. 14 h of audio - 60% in-domain(with as much incident/post-incident as possible) data; 40% out-of-domain; 70% of formal data; 30% of informal data; +/- 10% variance

Table 2: LoReHLT18 IL Data Description

7 Native Informant Resources

During the evaluation period, participants are allowed the use of a native informant (NI) in their system development. The LORELEI performers will be provided the native informant by their sponsor[9] through the data provider Appen. The native informant will be available remotely via telephone or internet connection. Open participants, if they wish to use a native informant, have to supply their own at their own cost and are free to determine how they communicate with their informant. However, consultation with the informant, by LORELEI performers and open participants, must abide by the following guidelines:

- Informant can be a native speaker of the IL but cannot be a professional linguist.
- It is up to the individual teams to determine how they will make use of the informant. However, **the evaluation data must remain unseen and sequestered, and all probings of the evaluation data are prohibited**. The teams must document how they have used the informant (e.g. producing additional resources for training, etc.).
- If a member(s) of the developer's team also happens to be a native speaker of the IL, this information must also be documented.
- Teams cannot ask the native informant about the incident regardless of the training conditions.
- For the constrained training condition, consultation with the informant is limited to the number of hours listed below for each IL and for each task a team participates regardless of how many submissions. If the use of the native informant exceeds the number of hours given, the submissions are considered to be in the unconstrained training track.
    - 1 h for Evaluation Checkpoint 1
    - 5 h for Evaluation Checkpoint 2 (4 h if 1 h was used in Checkpoint 1)

8 Evaluation Protocol

8.1 Evaluation Account

All evaluation activities will be conducted via an evaluation account. **There will be one account per team** so coordinate internally before you register. Go to https://goo.gl/forms/Xoj6mbLFU1V5BEJ93 to register for the evaluation. If you are not a LORELEI performer, you must sign the LDC data license agreement. The link to the agreement can be found in the registration form. The agreement is per site. If you have more than one site on your team, each site must sign the license separately. If everything is in order, an account will be created by NIST and a temporary password will be sent to the email provided in the

[9] LORELEI performers will be provided NI time by their sponsor only for the amount given above. If teams want additional time, they must make their own arrangement at their own cost.

registration form. We recommend that you change the password. You will make submissions from this account on behalf of your team.

8.2 System Input File Format

With the addition of the audio documents, LoReHLT18 has two input source formats.

8.2.1 Input Text Source Format

The input text source data for the MT, EDL, and SF tasks follows the LDC LTF common data format that conforms to the LTF DTD referenced inside the test files. An example LTF file is given below.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE LCTL_TEXT SYSTEM "ltf.v1.5.dtd">
<LCTL_TEXT>
  <DOC id="NW_ARX_UZB_164780_20140900" tokenization="tokenization_parameters.v2.0" grammar="none"
raw_text_char_length="1781" raw_text_md5="1511bf44675b0256adc190a7b96e14bd">
    <TEXT>
      <SEG id="segment-0" start_char="0" end_char="31">
        <ORIGINAL_TEXT>Emlashni birinchi kim boshlagan?</ORIGINAL_TEXT>
        <TOKEN id="token-0-0" pos="word" morph="none" start_char="0" end_char="7">Emlashni</TOKEN>
        <TOKEN id="token-0-1" pos="word" morph="none" start_char="9" end_char="16">birinchi</TOKEN>
        <TOKEN id="token-0-2" pos="word" morph="none" start_char="18" end_char="20">kim</TOKEN>
        <TOKEN id="token-0-3" pos="word" morph="none" start_char="22" end_char="30">boshlagan</TOKEN>
        <TOKEN id="token-0-4" pos="punct" morph="none" start_char="31" end_char="31">?</TOKEN>
      </SEG>
      <SEG id="segment-1" start_char="33" end_char="61">
        <ORIGINAL_TEXT>Pereyti k: navigatsiya, poisk</ORIGINAL_TEXT>
        <TOKEN id="token-1-0" pos="word" morph="none" start_char="33" end_char="39">Pereyti</TOKEN>
        <TOKEN id="token-1-1" pos="word" morph="none" start_char="41" end_char="41">k</TOKEN>
        <TOKEN id="token-1-2" pos="punct" morph="none" start_char="42" end_char="42">:</TOKEN>
        <TOKEN id="token-1-3" pos="word" morph="none" start_char="44" end_char="54">navigatsiya</TOKEN>
        <TOKEN id="token-1-4" pos="punct" morph="none" start_char="55" end_char="55">,</TOKEN>
        <TOKEN id="token-1-5" pos="word" morph="none" start_char="57" end_char="61">poisk</TOKEN>
      </SEG>
      ...
    </TEXT>
  </DOC>
</LCTL_TEXT>
```

8.2.2 Input Audio Source Format

The input audio source data for the SF task is a collection of segmented audio files in the .flac format.

8.3 System Output File Format

Each task has its own output format. Refer to the task specific section for information about the output requirement for that task.

8.4 Submission Requirements

All teams (LORELEI performers and open participants) are required to participate in the constrained training condition and are encouraged to participate in the unconstrained training condition. LORELEI performers are also required to submit at least one complete ensemble under the constrained training condition for each IL. Open participants are not required to have a complete ensemble. An **ensemble** is defined to be a set of submissions, one at each checkpoint, that developers of the system deem comparable over time. If a connection between checkpoints 1 and 2 cannot be made, LORELEI performers must perform an ablation study to provide information regarding how their systems behave under different factors (data, algorithm, time). Details of the ablation study are still in discussion.

Up to 10 ensembles can be created (10 rows with each row 2 slots one for each checkpoint). When a checkpoint is active, teams can upload their submissions to that checkpoint. **For a submission in the next checkpoint to belong to the same ensemble (the same row), that submission must be deemed comparable. Teams will document why they think they are comparable in their system description.** There will be no rearranging of the submissions into ensembles as in LoReHLT17.

Submissions will not be classified as primary or contrastive in LoReHLT18. For cross-team comparison, NIST will use the best scoring submissions at each given checkpoint regardless if they are from the same ensemble. Unlike LoReHLT17 no feedback will be given for any portion of the data. The only time replacing an existing submission is allowed is when it is determined the submission has a bug, at which time, teams will need to contact NIST to enable resubmission. Submissions that do not pass validation will not count toward the 10 submission limit.

At each submission, teams are recommended to provide a short description of their submissions when they upload their system output. At the conclusion of the evaluation, all teams are required to submit a more formal system description that covers their submissions for all tasks the team are participating in. The final results will be released to teams who submit a system description. Teams can download the template for the system description on the NIST LoReHLT website.

Refer to the task specific sections below for the requirements on how to package the system output for a given task into a submission file.

9 Evaluation Rules and Requirements

The evaluation is an open evaluation where the test data is sent to the participants who will process and submit the output to NIST. As such, the participants have agreed to process the data in accordance with the following rules:

- The participant agrees not to investigate the evaluation data. Both human/manual and automatic probing of the evaluation data is prohibited to ensure that all participating systems have the same amount of information on the evaluation data.
- The participant agrees to abide by the terms guiding the use of the native informant[10].
- The participant agrees to process at least the constrained training track for each of the selected tasks.
- The participant who is LORELEI performer agrees to complete all checkpoints to be considered a complete submission for each selected task and training track combination.
- The participant agrees to attend a post-evaluation workshop to present and discuss his/her systems.
- The participant agrees to the rules governing the publication of the results.

10 Guidelines for Publication of Results

This evaluation follows an open model to promote knowledge exchange with the wider community. At the conclusion of the evaluation cycle, NIST will create a report that documents the evaluation. The report will be posted on the NIST web space and will identify the participants and the scores from various metrics achieved for task.

[10] Contact NIST at lorehlt_poc@nist.gov if this presents a problem.

The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

10.1 Rules Governing Publication of Evaluation Results

The rules governing the publication of the LoReHLT evaluation results are similar to those used in other NIST evaluations.

- Participants are free to publish results for their own system, but participants must not publicly compare their results with other participants (ranking, score differences, etc.) without explicit written consent from the other participants.
- While participants may report their own results, participants may not make advertising claims about winning the evaluation or claim NIST endorsement of their system(s). Per U.S. Code of Federal Regulations (15 C.F.R. § 200.113): *NIST does not approve, recommend, or endorse any proprietary product or proprietary material. No reference shall be made to NIST, or to reports or results furnished by NIST in any advertising or sales promotion which would indicate or imply that NIST approves, recommends, or endorses any proprietary product or proprietary material, or which has as its purpose an intent to cause directly or indirectly the advertised product to be used or purchased because of NIST test reports or results.*
- All publications must contain the following NIST disclaimer:

  NIST serves to coordinate the evaluations in order to support research and to help advance the state- of-the-art. NIST evaluations are not viewed as a competition, and such results reported by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government.

11 Dry Run

The purpose of the dry run is to exercise the evaluation infrastructure, not testing systems' performance. As such, the dry run intends to be flexible and at the same time to follow the protocol of the official evaluation. Differences between the dry run and the official evaluation include:

- Shorter time duration between checkpoints
- No native informant
- The identity of the language is known before the IL Announcement (Mandarin, the same dataset used for the LoReHLT16 dry run)
- No scores will be reported. A feedback message will be presented to indicate if the submission has succeeded or failed. Sometimes detailed information on the nature of the failure may be provided.

Participants who are new to LoReHLT evaluation are encouraged to participate in a dry run evaluation to demonstrate evaluation readiness. Due to some changes in the protocol, previous LoReHLT participants are encouraged to participate in the dry run as well.

12 Uyghur Retest (LORELEI Performers Only)

LORELEI performers are required to reprocess the LoReHLT16 evaluation test set for the two tasks (MT and NER[11]). The goal of the retest is to show improvement/effect within teams in terms of novel approaches to language independent techniques and novel uses of information obtained from native informant. In effect, the retest is like checkpoint 5 *but* with no new data resources. Teams can use only sets 0, 1, S, 2, and data collected from NI from 2016 and can prepare these data in advance. During the retest (24 h), teams use their prepared components to process the evaluation set. Teams can also use data gathered from the extra 1 h they will have with the native informant during the retest. Below are some parameters regarding the retest:

- LORELEI performers should NOT use Set E Uyghur unsequestered portion for tuning or training but as an internal test set to test cross-language methods. Performers may use this unsequestered portion as training data for the official LoReHLT18 evaluation.
- LORELEI performers may NOT collect Uyghur-specific resources before or during the retest.
- LORELEI performers may use a non-Uyghur speaker to perform annotation during the retest.
- LORELEI performers may develop and use Uyghur-specific processing capabilities during the retest.
- LORELEI performers have 24 h to process the test data and submit the results. There is no checkpoint, but performers may make as many submissions as they wish. Feedback scores are provided for each submission.
- LORELEI performers will be provided some time with a native informant. Each team will have up to 1 h with the native informant per task. No additional time with the native informant is allowed before or during the retest, even at the performers' cost.
- LORELEI performers will inform NIST which submission NIST will report official results on.

13 Open Test on Additional Languages (OTAL)

In addition to the official test on two new ILs and the Uyghur retest, an open and optional set of tests on previously released LoReHLT incident languages are being offered for MT and EDL. This task removes many complexities of the main LoReHLT evaluation to focus on a common system across several languages and to lower the entry bar for non-LORELEI performers. All participants may choose to participate in either MT or EDL and in any number of the languages offered.

The IDs of the languages under test are known ahead of time. In fact, we will reuse the two languages from LoReHLT17 evaluation (Oromo and Tigrinya). LORELEI performers must take special care to remove the data from their systems for participation in this task. Participants will be given basic incident language training data (Set 0, Set 1, Set 2, Set S) as soon as they register for the evaluation. Participants may use all publicly available additional data for training as long as this data do not fall between the blackout periods. At the start of the evaluation participants will receive the evaluation data and have 3 days to process and return the output to NIST for scoring. It is a single-checkpoint evaluation. No native

[11] NER task definition can be found in the LoReHLT16 evaluation plan at https://www.nist.gov/itl/iad/mig/lorehlt16-evaluations

informant is allowed. NIST will not provide any score feedback. Teams can submit up to 10 submissions. The input source and the system output for this task follow the same format as the official test.

14 Schedule

Milestone	***Date***
Initial version of evaluation plan published	Jan 15, 2018
Registration period	Mar 01 – May 31, 2018
6-month PI meeting (LORELEI performers only)	Mar 20 – 22, 2018
Uyghur Retest	May 2018
Dry Run	May 2018
LoReHLT18 Evaluation	Jul 2018
OTAL Evaluation	Sep 2018
DARPA PI meeting (LORELEI performers only)	TBD
Human Assessment	TBD
NIST post-evaluation workshop co-located with TAC/TREC (pending number of participants)	TBD
Uyghur Retest Schedule (LORELEI Performers Only)	
Evaluation data available[12]	12:00 noon EDT May 09
System output submission for retest due	12:00 noon EDT May 10
Dry Run Schedule	
Encrypted data released by LDC	May 14
IL Announcement - Decryption keys for set 0 and set E distributed	12:00 noon EDT May 15
Evaluation Checkpoint 1 - System description submission opens - System output submission for Evaluation Checkpoint 1 opens - Decryption key for set 1 and set S distributed at end of Evaluation Checkpoint 1 and after system output submission made	12:15pm EDT May 15 – 12:00 noon EDT May 16
Evaluation Checkpoint 2 - System output submission for Evaluation Checkpoint 2 opens	12:15pm EDT May 16 – 12:00 noon EDT May 17
System description submission closes	12:15pm EDT May 17
Preliminary results released if system description is received	May 18
LoReHLT18 Evaluation Schedule	
Encrypted data released by LDC	Jun 29
IL Announcement - Decryption keys for set 0 and set E distributed	12:00 noon EDT Jul 02
Evaluation Checkpoint 1 - System description submission opens - Access to Native Informant (MT, EDL, SF; see below) - System output submission for Evaluation Checkpoint 1 opens - Decryption key for set 1 and set S distributed at end of Evaluation Checkpoint 1 and after system output submission made	12:15pm EDT Jul 02 – 12:00 noon EDT Jul 03
Evaluation Checkpoint 2	12:15pm EDT Jul 03

[12] LORELEI performers should have the evaluation data already.

- Access to Native Informant (MT, EDL, SF; see below) - System output submission for Evaluation Checkpoint 2 opens	– 12:00 noon EDT Jul 11
System description submission closes[13]	12:00 noon EDT Jul 20
System description reviewed by NIST	Jul 23
Preliminary results released if system description is received	Jul 24
Native Informant Timeline (time amount is per incident language per team per task (MT, EDL, SF)	
Up to 1 h between 12:15 ET Jul 02 to 12:00 ET Jul 03 Up to 5 h between 12:15 ET Jul 03 to 12:00 ET Jul 11 (or 4 h if 1 h was used between Jul 02 and Jul 03)	
OTAL Evaluation Schedule	
Encrypted data released by LDC	Sep 14
Decryption keys distributed by NIST	Sep 17
System submissions due to NIST	Sep 21

15 Machine Translation (MT) Evaluation Specifications

15.1 Task Definition

Given a text document in the incident language, the MT system is required to automatically translate the document's content into English. The entire IL only portion of the test set must be translated, even though only a subset of it will be scored in the machine translation evaluation. MT systems are to ignore the English portion of the test set which includes not to process, not to probe, and not to inspect the data, as outlined in the evaluation rules and requirements in section 9 Evaluation Rules and Requirements.

15.2 Performance Measurement

The goal for the assessment of the MT output is to evaluate it in the context of the larger LORELEI task. Several different approaches, outlined below, will be implemented to achieve this. In addition, NIST will continue to investigate additional automatic approaches geared towards measurement of successful translation of content relevant to the LORELEI task.

Some of the measurements described below will be carried out on subsets of the MT test set based on annotation by the SF systems. SF systems will be required to identify exactly one segment (using the the segmentation provided for MT processing) for each document and detected SF. These segments are likely of higher relevance to the LORELEI task. Measuring MT performance on only these may provide better insight for assessing the impact of MT on the LORELEI task. For the scoring and annotations described below, this subset will then be reduced to only those instances where the SF system identified the SF correctly, and naturally to only those instances that are part of the MT test set as well. The exact details of the protocol for this are still being finalized.

[13] While we ask that each team produces one system description for all tasks, if your team participates in SF Speech which has a later system description deadline, we ask that you resubmit the system description with the SF Speech info added so you will get your text results at the earlier result release date.

15.2.1 Evaluation of Impact of Machine Translation on Situation Frame Performance

In order to assess the degree to which MT aids SF performance, SF scoring will be performed on the reference translation and selected MT outputs in addition to SF scoring on the source. This will naturally be limited to those SF systems that have the capability of processing English translations, not just the source data directly. This will allow for a comparison of SF performance on:

- the source data,
- the English reference translation, and
- the English MT output.

Additionally, it will allow for a correlational analysis of automatic SF and MT scores on the same English MT output (see section 15.2.2 Automatic MT Metric Scoring below for automatic MT scores that will be computed).

15.2.2 Automatic MT Metric Scoring

The MT output will also be scored with fully automatic MT metrics, to include METEOR and potentially others. Scoring will be done using a single reference translation. Case will be preserved. Normalizations may be implemented for scoring purposes as necessary for the domains and data encountered, such as preventing URLs from being tokenized into multiple pieces.

Scoring will be performed separately for different portions of the MT subset of set E:

1. The entire MT subset of set E, with scores at the system, document, and segment levels,
2. The subset of SF justification segments described above.

15.2.3 Human MT Assessment

An additional human assessment step may be performed, in which assessors will judge MT output on the subset of SF justification segments (and potentially surrounding segments) as to whether the MT would allow for the identification of the correct situation frame. The exact details of the protocol for this are still being finalized.

15.3 MT System Output Format

MT systems are required to output the translation conforming to the lorehlt-mt-v1.2.dtd[14]. A sample MT system translation file is given below:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE mteval SYSTEM "lorehlt-mt-v1.2.dtd">
<mteval>
  <tstset>
    <doc docid="NW_ARX_UZB_164780_20140900">
      <seg id="segment-0"> Who did vaccinations first?</seg>
      <seg id="segment-1"> Go to navigation, search</seg>
      …
    </doc>
  </tstset>
</mteval>
```

[14] ftp://jaguar.ncsl.nist.gov/lorehlt16/lorehlt-mt-v1.2.dtd

The value of each `doc docid` attribute or `seg id` attribute must match exactly to that used in the original LTF file.

Note that there is one MT system output file for each MT system input file, and the output file must have the same name as the input file.

15.4 System Submission Format

The MT system output files as described in 15.3 MT System Output Format should be placed into flat-file hierarchy and compressed into a .tgz or .zip file. There are no restrictions on the submission file name besides the suffix '`.tgz`' or '`.zip`'.

16 Situation Frame (SF) Evaluation Specifications

16.1 Task Definition

Given a collection of text and speech documents, in the incident language and/or English, an SF system is required to automatically identify zero or more situation frames covered in the document and build a knowledge base (KB) of situations by identifying situation frames for a particular situation type and place (geographic location) within each document. The combination of type and place uniquely identifies a situation. Additional attributes may be included with the situation frame or situation. For the place field, the participants are to choose a place name from a comprehensive repository of place names (e.g. GeoNames) that the LDC will provide. For scoring purposes, the place name should match precisely the reference, and partial credit will not be given for partial overlap or containment. For instance, if the reference points to a repository entry for "Reston, VA" and the SF system reports repository entry of "Fairfax County, Virginia", the place field of the frame will be considered labeled wrong despite "Reston" being in "Fairfax County" because the annotator was able to determine the location more precisely from the source document, and the SF system is expected to do the same.

A **document-level situation frame** has the following **required** structure (text-only*; need-type-only†):

- **DocumentID**: The document from which the SF was extracted
- **Type**: The situation type, from a fixed set of needs / issues. One of "evac", "food", "infra", "med", "search", "shelter", "utils", or "water" for a need frame. Or, one of "regimechange", "crimeviolence", or "terrorism" for an issue frame.
- **Place_KB_ID**: The KB ID of the location at which the situation is/was present, from the KB provided by LDC. In the event that the system is confident that no place should be associated with the frame, the system is expected to return an empty string in this field.
- **Status**: The temporal need or issue status of the situation. For "need" frame either "past", "current", or "future"; for "issue" frame either "current" or "not_current".
- **Confidence**: How confident the system is in the creation of the situation frame from the document, ranging from 0 to 1, inclusive.
- **Justification_ID***: The segment ID of one segment from the source document justifying the creation of the situation frame. This field is required for LORELEI performers but is optional for open participants[15]. Please note that the "Justification" field this year is used for human assessment purposes for MT (see 15.2.3 Human MT Assessment), and NIST may use this information for various exploratory measurements.
- **Resolution†**: Either "sufficient" or "insufficient" (also if not known to be sufficient, considered "insufficient").
- **Urgent**: Either "true" or "false" (note, this field is required for both "need" and "issue" frames)

The LDC will create a gold standard collection of document-level situation frames, which will be aggregated prior to scoring to create a list of KB-level situations.

The entire test set must be processed even though only a subset of documents will be scored in the SF evaluation. Systems must provide at least the **DocumentID**, **Type**, **Place_KB_ID**, **Status**, **Resolution**, and

[15] This year, human assessment will only focus on the text documents and the Justification field for the speech documents will be ignored, if reported.

Urgent fields for each situation frame in order to be evaluated. The diagnostic metrics also require the **Confidence** field to be meaningful.

16.2 Performance Measurements

The conceptual use of SF technology is to support downstream applications that aggregate SF outputs to provide situational awareness using a variety of data sources that differ substantially with respect to the density of SFs and that simultaneously provides detailed supporting information about the situation. Thus, systems must directly support both low miss and low false alarm application scenarios as well as provide high quality supporting information.

This year's SF evaluation will start addressing the aggregation test case by introducing the concept of "KB-level situations". During the scoring phase, the situation frames from each system submission will be labeled for gravity and grouped into KB-level situations as described further in this section. NIST will focus on evaluating (1) correct identification of the KB-level situations and (2) inference of gravity of KB-level situations.

The primary SF system performance metric is nDCG. The rationale for the choice of nDCG as the primary metric is centered around the needs of the analyst. From the perspective of the T&E team there needs to be a single primary metric for which the participants can optimize their systems. This metric should represent as much as possible the desired needs of the analyst that will be using the system. For situational awareness the analyst would be focusing on different situations that require analyst's attention. Therefore for this year the evaluation focuses on KB level situations instead of single situation frames.

The system would provide the analyst a list of situations. While it is desired for the list to be of good precision and good recall, it is more important not to miss urgent situations that are ongoing and require immediate attention than those that are less urgent, or have already been resolved. A cumulative gain (CG) metric addresses this requirement and gives higher gain to more urgent situations that are correctly identified than less urgent situations or situations that are not current anymore. Additionally, there is going to be potentially many more situations than the analyst can focus on at once. Therefore it is important to present the analyst the list of situations in an order such that higher priority situations appear closer to the top of the list. To address this stipulation, a cumulative gain needs to be discounted based on the position of the situation in the list. In order to compare performance of the system across multiple lists of situations we normalize the final score and arrive at an nDCG metric.

To determine which situations are more important for the purpose of this evaluation, a notion of gravity is introduced. A situation frame is considered grave if it is "current", "urgent", and "unresolved". The number of grave situation frames in a KB-level situation is meant to indicate the magnitude and seriousness of the situation. Therefore, the gain in the nDCG metric will be assigned based on the number of grave situations and will be determined once the situation frame annotations become available. For illustrative purposes, a KB-level situation with 25 or more grave situation frames could be assigned a "High" gain, 10-25 grave situation frames assigned a "Medium" gain, and less than 10 assigned a "Low" gain, where gain might be set to 5 for a "High" KB-level situation gravity, 3 for a "Medium" situation gravity and 1 for a "Low" situation gravity.

Note that the systems will get credit for all correctly identified situations, not just current, urgent & unresolved. If a system occasionally mislabels fields, but the KB-level situations still end up in the right order, the system could still get the highest possible nDCG score. The score decreases with false alarms (KB-level situations that don't exist), and/or by listing less grave situations before more grave ones.

The nDCG metric alone might be too broad to be indicative of areas where the systems fail, and what system components need to be improved, but since the participants' systems are complex and vastly differ in architecture, a narrower performance metric that is useful for one participant might not be useful for another, and does not provide a part-by-part comparison of systems. In order to help facilitate teams' R&D efforts NIST will provide a variety of other metrics for diagnostic purposes during the analysis phase. Two types of diagnostic metrics will be reported to evaluate system performance with respect to finding all the correct situation frames: macro-averaging metrics to evaluate systems' ability to correctly identify situations through detection and clustering of document-level situation frames, and KB-level situation metrics to evaluate systems' ability to infer situation gravity.

Detected situation frames will be linked to KB-level situations by "type" and "place"; and as a simplifying assumption this year, participants should consider all situation frames with a common type and location to refer to the same KB-level situation.

This year, multiple references will be used for scoring as well. The Precision and Recall metrics in this section are short for Occurrence Weighted Precision and Occurrence Weighted Recall. The weights for each frame are determined by the number of occurrences in the combined reference with respect to equivalence class. False positives are given a weight of 1 for the purposes of computing Occurrence Weighted Precision.

16.2.1 Annotation of Urgency and Its Interpretation for Scoring Purposes

Due to poor inter-annotator agreement for urgency decision on Situation Frames, for Y3 LDC will use a new approach designed for better consistency. Annotators will label two features: "Severity" and "Scope". Severity indicates the most severe likely outcome based on what is/was expressed in the document, and scope indicates the highest number of people potentially affected. The two metrics are considered orthogonal and each can assume four possible values. Scope can be of "Individual/ Small Group", "Large Group", "Municipality", "Multiple Municipalities"; severity can be of "Inconvenience/Discomfort", "Non-life Threatening Injury or Destruction", "Possible Loss of Life", "Certain Loss of Life". Since the SF participants continue this year to label "Urgent" frames with a binary label, the "Scope" and "Severity" annotations are combined and converted to the single binary value for scoring purposes as follows: if a situation frame is of at least the scope of a "Large Group", or severity of at least "Non-life Threatening Injury or Destruction", the frame will be considered urgent.

16.2.2 Primary Metric: Normalized Discounted Cumulative Gain Metric

To evaluate systems' ability to infer situation gravity at the KB-level, Normalized Discounted Cumulative Gain (nDCG) metric will be computed. We consider the gravity of a situation to be the number of constituent situation frames that are current, urgent, and unresolved for a particular KB-level situation. nDCG uses a graded relevance scale of KB-level situations in the result set, measuring the gain (usefulness) of a given KB-level situation based on its position in the result list. Situations are binned by gravity and assigned gains based on the bin (e.g. "High": 5 points, "Mid": 3 points, "Low": 1 point). For scoring purposes the focus is on systems' ability to correctly order situation frames "High" before "Mid" before "Low", and the complete ordering is not important. Thus, the gain is accumulated from the top of the result list to the bottom and the gain of each result discounted at lower ranks and then normalized.

Normalized Discounted Cumulative Gain is defined as follows:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

where DCG_p is Discounted Cumulative Gain at rank p (the first p gravest KB-level situations) and is defined as:

$$DCG_p = \sum_{i=1}^{p} \frac{Gain_i}{log_2(i+1)}$$

where $Gain_i$ is the gain value of KB-level situation i.

Ideal Discounted Cumulative Gain (IDCG) is the best possible DCG. It is used as a denominator to normalize the DCG of the system and is calculated by applying the DCG_p formula above to the sorted reference list of KB-level situations.

16.2.3 Diagnostic Metrics

Diagnostic metrics over KB-level situations will be evaluated and scored using the notion of equivalence classes of:

- *type, place*
- *type, place, status*
- *type, place, status, relief*
- *type, place, status, urgency*
- *type, place, status, relief, urgency*
- *type, place, status, relief, urgency* (filtering only for urgent and unresolved frames)[16]

These equivalence classes determine what it means for a given SF frame to be relevant in each KB level situation. Reported metrics include "Mean Average Precision" and "Macro-Average Recall". In Average Precision each correctly detected SF is counted as "relevant", and we consider the ranking of SFs ordered by confidence. The gold standard comprises the known relevant frames. "Mean Average Precision" is the overall metric of all Average Precision metrics averaged across all KB-level situations, and "Macro-Average Recall" is the overall recall metric averaged across all KB-level situations.

Mean Average Precision is defined as follows:

$$Mean\ Average\ Precision = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

where

Q - is the total number of KB-level situations

$AP(q)$ - is the average precision of KB-level situation q and is defined as:

$$AP = \frac{\sum_{k=1}^{n} (Precision(k) \times rel(k))}{\#\ of\ relevant\ docs}$$

$rel(k)$ equals 1 if k is a relevant situation frame, 0 otherwise

Macro-Average Recall is the average of Recall measures across all KB-level situations and is defined as follows:

[16] To measure the finer grain system performance with respect to gravity, the same macro-averaged precision/recall measures as described above will be used but only considering the set of "urgent", "unresolved" situation frames for a given "current" situation as relevant.

$$Macro - Average\ Recall = \frac{1}{n}\sum_{j=1}^{n} R_j$$

where

n is the number of KB-level situations

R_j is the recall measure of KB-level situation j and is defined as

$$R = \frac{TP}{TP+FN}$$

16.2.4 Diagnostic KB-level Precision at k Metric

Another KB-level situation metric is Precision at N, where N is a number of situations above a certain gravity threshold in the "High" category. The KB-level situations generated by participants' systems will be sorted by gravity and the top N situations will be compared to the sorted reference list of KB-level situations using the precision metric as follows:

$Precision(k)$, precision at cut-off k, defined as:

$$Precision(k) = \frac{TP(k)}{TP(k)+FP(k)}$$

16.2.5 Diagnostic Precision-Recall Curves

For each system submission and for each equivalence class a Precision-Recall (PR) curve will be generated, with each point of the curve corresponding to a recall on the x axis and precision on the y axis. The curve will be produced by ordering the situation frames by confidence in descending order and sweeping across the confidence values in the system output calculating precision and recall at each situation frame. Additionally, the plot if each curve will include the Area Under the Curve (AUC) as an aggregate metric.

16.3 Scoring Procedure

This section uses a high-level pseudocode to describe the steps in the scoring process. Please note that some loops can be folded for efficiency in the scorer, but are repeated below to provide better clarity of the scoring procedure. See Appendix A - SF Scoring Example.

First, we group reference situation frames and system situation frames into subsets of unique "type" and "place" that represent KB-level situations. Then, we sort each system subset in descending order using the "Confidence" field.

Then, for each equivalence class described in section 16.2.1 Diagnostic Metrics we compare the set of system output frames against the set of reference frames to compute the metrics:

Compute primary metric:

1. Normalized Discounted Cumulative Gain:
    1.1. Count the number of grave situation frames in each reference KB-level situation

1.1.1. Based on Gravity, assign gain[17] value to each reference KB-level Situation
1.1.2. Sort the reference KB-level situations by gain in descending order
1.1.3. Compute Ideal Discounted Cumulative Gain (IDCG) for normalization purposes

1.2. Count the number of grave situation frames in each KB-level situation that the participant system reported
1.2.1. Order KB-level situations in descending order by number of Grave situation frames
1.2.2. Compute Discounted Cumulative Gain

1.3. Normalize the Discounted Cumulative Gain from previous step, using IDCG

2. Precision at N:
2.1. Sort reference KB-level situations in descending order by number of grave situation frames
2.2. Sort KB-level situations that the participant system reported in descending order by number of grave situation frames
2.3. For N from 1 to the total number of KB-level situations
2.3.1. Compute Precision of subset of first N KB-level situations

Compute diagnostic metrics, for each **equivalence class**:

3. Mean Average Precision:
3.1. For each reference KB-level situation Q:
3.1.1. Find the matching "type, place" system KB-level situation
3.1.2. Compute Average Precision of each system KB-level situation, ordering Situation Frames by "Confidence"
3.2. Compute the mean of the Average Precisions from the previous step over all KB-level situations

4. Macro-Average Recall
4.1. For each reference KB-level situation Q:
4.1.1. Find the matching "type, place" system KB-level situation
4.1.2. Compute the Recall of each KB-level situation
4.1.3. Compute the average of all Recalls from previous step

Compute situation frame metrics:

5. PR Curves:
5.1. Remove all frames below the current confidence threshold
5.2. Transform the remaining frames to the current equivalence class
5.3. Calculate True Positives, False Positives and False Negatives taking into account the fields of the current equivalence class
5.4. Calculate Precision and Recall

[17] Each situation will be assigned "gain" value representing High, Medium, or Low gain based on the number of grave situation frames in the KB-level situation in question. The range of grave situation frames that corresponds with each gain level, as well as the gain values will be determined by NIST at a later stage of the evaluation after analyzing the annotated datasets from LDC for this year's evaluation, once they become available.

16.4 System Output Format

The system output structure is a JSON structure and should conform to the json schema. The latest schema (version 2, filename: “LoReHLT18-schema_V2.json”) along with the latest LoReHLT Frame Scorer software package (LoReHLT18_SF_Scorer_0.9) can be downloaded from the official LoReHLT ‘18 webpage. Note that the schema version and the scorer version will be different (updated) from the last year’s evaluation and will be made available as soon as they are ready. Contained below is a simple example of the system output structure for this year’s SF task.


```
[
 { "DocumentID": "CMN_NG_000031_20080707_80020000G",
   "Type": "infra",
   "Place_KB_ID": "KBID099324",
   "Status": "current",
   "Confidence": 0.4,
   "Justification": "segment-5",
   "Resolution": "insufficient",
   "Urgent": false},

 { "DocumentID": "CMN_NG_000031_20080707_80020000G",
   "Type": "shelter",
   "Place_KB_ID": "KBID085430",
   "Status": "not_current",
   "Confidence": 0.6,
   "Justification": "segment-7",
   "Resolution": "insufficient",
   "Urgent": false}
]
```


16.5 System Submission Format

The SF system output files as described in section 16.4 System Output Format named “system_output.json”. There are no restrictions on the submission file name besides the suffix “`.tgz`” or “`.zip`”.

17 Entity Discovery and Linking (EDL) Evaluation Specifications

17.1 Task Definition

Given a document collection in the incident language (IL) and English, an EDL system is required to automatically identify entity mentions, classify them into predefined entity types, and link them to a pre-assembled Knowledge Base (KB). In addition, for entity mentions that do not have KB entries, i.e. NIL entity mentions, an EDL system must cluster them.

For documents in IL, the mention type is still limited to named mentions as before, but for English documents, an EDL system must also discover and link nominal mentions. NIST will report scores with and without English nominal mentions.

The entity types continue to be Geo-Political Entity (GPE), Location (LOC) including Facility (FAC) as defined in other entity-related tasks, Person (PER), and Organization (ORG).

For more details on the NER part, please consult LDC's Simple Named Entity Annotation Guidelines. LDC has also released EDL annotation guidelines specifically tailored for LOREHLT. Both are available where LORELEI materials are stored. If you are an open participant and do not have direct access to the web site, please contact LDC at lorelei-poc@ldc.upenn.edu.

Participants may also refer to TAC KBP 2016 for EDL annotation guidelines, a copy of which can be accessed at: https://tac.nist.gov/2016/KBP/guidelines/TAC_KBP_2016_EDL_Guidelines_V1.1.pdf

17.2 Knowledge Base (KB)

The reference KB – all in English and one each IL – will consist of four input sources as follows. For details, please refer to the relevant document released by LDC.

1. GeoNames (http://www.geonames.org/) for GPE and LOC entities;
2. CIA World Leaders List (https://www.cia.gov/library/publications/world-leaders-1/) for PER entities;
3. Appendix B of the CIA World Factbook for ORG entities https://www.cia.gov/library/publications/resources/the-world-factbook/appendix/appendix-b.html ;
4. Manually augmented incident-, region- and/or domain-relevant PER and ORG entities that do not appear in (1) through (3).

A small sample KB will be distributed before evaluation so that new participants may become familiar with the format. The sample KB will include a few examples of manually augmented entries, unrelated to any IL's to avoid exposing evaluation-sensitive information.

17.3 Performance Measurements

Scoring metrics from the TAC KBP 2016/2017 EDL task will be extended to the EDL task. Specifically, Precision, Recall and F1 scores will be reported for the following metrics:

Mention Evaluation

- strong_mention_match (NER)
- strong_typed_mentin_match (NERC)
- overlap_maxmax_micro
- overlap_maxsum_micro
- overlap_summax_micro
- overlap_sumsum_micro

Linking Evaluation

- strong_typed_all_match (NERLC)
- strong_typed_link_match (NELC)
- strong_typed_nil_match (NENC)

Tagging Evaluation

- entity_match (KBIDs)

Clustering Evaluation

- Mention_ceaf (CEAFm)
- Typed_mention_ceaf (CEAFmC)
- Typed_mention_ceaf_plus (CEAFmC+)

Clustering Diagnostics

- mention_ceaf;docid=<micro> (CEAFm-doc)
- mention_ceaf:is_first:span (CEAFm-1st)

For more details on these metrics, refer to section 2.2 in the 2015 KBP overview paper at http://nlp.cs.rpi.edu/paper/kbp2016.pdf and section 14.2 in the 2016 LoReHLT evaluation plan at https://www.nist.gov/file/326366.

The EDL scorer is posted at https://github.com/wikilinks/neleval.

17.4 System Output Format

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. Please note that while the format is identical to that of TAC 2015/2016 EDL.

```
Field1<tab>Field2<tab>Field3<tab>...<tab>Field8
```

where:

Field 1: system run ID, unique team_id to identify each team and their runs

Field 2: mention ID, unique for each entity name mention

Field 3: mention head string, the full head string of the 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: a KB link entity ID (numeric) or NIL clustering ID (NIL followed by a sequence of digits)

Field 6: entity type: {GPE, ORG, PER, LOC} type indicator for the entity

Field 7: mention type {NAM, NOM}

Field 8: a confidence value, 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

Sample EDL output:

```
NIST    QUERY300 Singapore         ENG_DF_001503_20070729_G00A0AFCA:889-897   m.06t2t  GPE     NAM     1.0
NIST    QUERY301 Singapore         ENG_DF_001503_20070729_G00A0AFCA:1048-1056 m.06t2t  GPE     NAM     1.0
NIST    QUERY303 Jollytinker       ENG_DF_001503_20070729_G00A0AFCA:1620-1630 NIL45    PER     NAM     1.0
NIST    QUERY304 Asia              ENG_DF_001503_20070729_G00A0AFCA:1344-1347 m.0j0k   LOC     NAM     1.0
```

Each system submission will be validated to ensure it conforms to the specifications. If validation fails, it will be rejected and will not be scored. The validation script is available at the LORELEI website. If you are an open participant and cannot retrieve it from the website, please contact NIST for a copy.

17.5 System Submission Format

Each aforementioned EDL output file, preferably with the .tab extension, should be packaged into a single flat tarball with an extension of either .tgz or .tar.gz, and each submission must have be uniquely named. The submission file name should include information about the team's identity, task, checkpoint, and run id, etc., for example, NIST_EDL_CP1_1.tab.tgz (which would be unzipped as NIST_EDL_CP1_1.tab).

18 SF Scoring Example

This appendix provides examples to better convey how the SF Task metrics are being computed by the scorer.

18.1 Primary Metric: Normalized Discounted Cumulative Gain Example

For the sake of this example, assume that the gain is 5 for "High", 3 for "Medium", and 1 for "Low". Further assume that a KB-level situation with 25 or more grave situation frames is assigned a "High" gain, 10-25 grave situation frames is assigned a "Medium" gain, and less than 10 is assigned a "Low" gain.

KB-level Situation	**# of grave SF's**	**Gain**	**Rank p**
Sit A	100	5	1
Sit B	30	5	2
Sit C	26	5	3
Sit D	24	3	4
Sit E	19	3	5
Sit F	11	3	6
Sit G	5	1	7
Sit H	3	1	8
Sit I	2	1	9
...	...	...	...
Sit Z	0	0	p

Further suppose a system generated situation frames that resulted in the following list of 9 situations:

KB-level Situation	**# of grave SF's**	**Gain**	**Rank p**
Sit A	100	5	1
Sit D	29	3	2
Sit C	21	5	3
Sit E	19	3	4

Sit B	9	5	5
Sit F	7	3	6
Sit G	5	1	7
Sit H	3	1	8
Sit I	2	1	9

The DCG_p of this list of 9 situations is computed as follows:

$DCG_1 = \sum_{i=1}^{1} \frac{5}{log_2(1+1)} = 5$, $DCG_2 = \sum_{i=1}^{2} \frac{Gain_i}{log_2(i+1)} = 5 + \frac{1}{log_2(3)} = 6.89$,

$DCG_3 = \sum_{i=1}^{3} \frac{Gain_i}{log_2(i+1)} = 5.63 + \frac{3}{log_2(3)} = 9.39$ and so on.

$DCG_p = \{5,\ 6.89,\ 9.39,\ 10.68,\ 12.62,\ 13.69,\ 14.02,\ 14.34,\ 14.64\}$

The Ideal Discounted Cumulative Gain uses the same formula applied to the values from the reference table and results in:

$IDCG_p = \{5,\ 8.15,\ 10.65,\ 11.95,\ 13.11,\ 14.18,\ 14.51,\ 14.82,\ 15.13\}$

The normalized discounted cumulative gain is computed by dividing the discounted cumulative gain by the ideal discounted cumulative gain:

$nDCG_p = \frac{DCG_p}{IDCG_p} = \{1,\ 0.85,\ 0.88,\ 0.89,\ 0.96,\ 0.97,\ 0.97,\ 0.97,\ 0.97\}$

18.2 Diagnostic Metrics

Suppose for a given KB-level situation, the system reported the following situation frames:

Document ID	**Type**	**Place**	**Confidence**
SF1	food	Washington, DC	0.97
SF2	food	Washington, DC	0.92
SF5	food	Washington, DC	0.89
SF3	food	Washington, DC	0.87
SF4	food	Washington, DC	0.73

Note that the frames are sorted in descending order by Confidence.

Suppose the reference for this KB-level situation is as follows:

Document ID	Type	Place
SF3	food	Washington, DC
SF2	food	Washington, DC
SF1	food	Washington, DC
SF7	food	Washington, DC

Note that the order of situation frames in the reference KB-level situation does not matter.

18.2.1 Mean Average Precision Example

$Precision(1) = 1/1 = 1$, because SF1 (first SF in the system reported list) is in the reference list.

$Precision(2) = 2/2 = 1$, both SF1 and SF2 are in the reference.

$Precision(3) = 2/3$, SF5 is not in the reference.

$Precision(4) = 3/4$, SF1, SF2. SF3, are in reference, SF5 is not.

$Precision(5) = 3/5$, SF4 is also not in the reference.

For average precision calculation, SF1, SF2, SF3 get relevance of 1, and SF4, SF5 relevance of 0; and number of relevant documents is 4, because there are four documents in the reference list.

Thus:

$$Average\ Precision = \frac{(1/1)*1 + (2/2)*1 + (2/3)*0 + (3/4)*1 + (3/5)*0}{4} = \frac{2.75}{4} = 0.69$$

Suppose there were 5 KB-level situations, and on each of these situations a given system attained average precision of 0.69, 0.97, 0.84, 0.92, 0.78

$$Mean\ Average\ Precision = \frac{0.69 + 0.97 + 0.84 + 0.92 + 0.78}{5} = 0.84$$

18.2.2 Macro-Average Recall Example

For the KB-level situation example presented above, the system correctly identified situation frames SF1, SF2, SF3, and missed SF7. Therefore, the recall is 0.75

Suppose there were 5 KB-level situations, and on each of these situations a given system attained a recall of 0.75, 0.92, 0.61, 0.32, 0.66

$$Macro - Average\ Recall = \frac{0.75 + 0.92 + 0.61 + 0.32 + 0.66}{5} = 0.65$$

18.2.3 Precision at N Example

Following the example system output above, Precision at N for the gravest situations that were assigned "High" gain would be Presision at 3. The system correctly identified Sit A and Sit C, but missed Sit B, therefore:

$$Precision(3) = \frac{2}{3} = 0.66$$

Precision at N for the situations with 15 or more grave frames would be Precision at 5. The system correctly identified situations A, C, D, E, but missed B, therefore:

$$Precision(5) = \frac{4}{5} = 0.8$$