# NIST LoReHLT 2017 Evaluation Plan

Last Updated: July 27, 2017

Revision History

July 26, 2017 (v2.3.3):

- Added version numbers to DTD, schema, and annotation guidelines used by each task
- Clarified submission protocol

July 5, 2017 (v2.3.2):

- Extended SF Speech deadline to 12:00 PM ET August 31, 2017
- Revised system description deadline and results release date to accommodate SF Speech extension
- Clarified CP duration
- Added LORELEI PI meeting dates to schedule
- Added KB as a separate data resource

June 20, 2017 (v2.3.1):

- Checkpoint 3 for text now has 5 hours of NI
- Speech also has 5 more hours of NI
- Clarified Uyghur retest purpose
- Added ablation study

June 07, 2017 (v2.3):

- SF Speech now has only 1 checkpoint

# 1 Introduction

The 2017 LoReHLT evaluation is the second evaluation in the NIST Low Resource Human Language Technology evaluation series that began in 2016. The series was designed in collaboration with the 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 and track the progress made.

While the 2017 evaluation will be similar to the 2016 evaluation in many respects, the 2017 evaluation will include two surprise languages instead of one. The situation frame task will be extended to speech data. Additionally, there will be no distinction between primary or contrastive systems, and teams can submit up to 10 submissions per checkpoint and will be able to get score feedback on 10% of the dataset; and finally organization of submissions into ensembles will be done at the last checkpoint.

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[1].

# 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)** – for each document, automatically translate it from a given incident language (IL) to English. For MT specific requirements, see Section 14.
- **Situation Frame Text** – for each document, automatically generate Situation Frames covered in the document. For SF Text specific requirements, see Section 15.
- **Situation Frame Speech[2]** – for each audio recording, automatically generate Situation Frames covered in the recording. For SF Speech specific requirements, see Section 16.
- **Entity Discovery and Linking (EDL)[3]** – for each document, identify the named mentions, classify them into pre-defined entity types, and link the mentions to a knowledge base. For EDL specific requirements, see Section 17.

[1] http:/www.nist.gov/itl/iad/mig/lorehlt17-evaluations

[2] It is expected that both SF tasks (text and speech) will converge in future years, but for 2017 they are two separate tasks due to slightly different task definitions and data annotated to different guidelines.

[3] This task has evolved from a simple named entity recognition task (identifying and classifying named mentions as required in LoReHLT16) to include also linking these named mentions to a knowledge base.

# 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 about the incident after the incident is announced as both would constitute "knowing the future". In a live situation, information about the incident will develop over time, and systems will get to learn more about it. 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[4].

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[5].

# 4 Training Conditions

For each evaluation task, there are two training conditions (constrained and unconstrained) that differentiate the amount/source of incident language-related training material without preventing/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 Section 5. 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 evaluation 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, 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 is permitted. Prior to the incident and incident language announcement, teams can assemble

[4] Please see Section 7 for a complete guidelines regarding the native informant usage.

[5] 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.

mono- and bi-lingual resources so long as they do not include the incident language. 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). 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[6]. Prior to the incident and incident language announcement, teams can assemble mono- and bi-lingual 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 Text, SF Speech, 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 text, or entity discovery and linking guidelines and Appen (TBD) for the LoReHLT situation frame speech annotation guidelines.

# 6 Evaluation Data

The LoReHLT17 will have **two** incident languages which will be referred as IL5 and IL6. Each incident language follows the same data component and format as described below.

## 6.1 Component Definition & Release Plan

MT, EDL, and SF Text will be evaluated at all three checkpoints while SF Speech will be evaluated only at the last checkpoint. The LDC will release both the text and speech data for both ILs in an encrypted format (see Section 6.4) 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 (Aug 4)
  - **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 2**: Encrypted incident/post-incident IL training data set 2 released
  - **Set S**: Encrypted incident/post-incident English Scenario Model released
  - **Set E**: Encrypted incident/post-incident IL text evaluation data released
  - **Set 0 Speech**[7]: Encrypted IL training data released
  - **Set E Speech:** Encrypted IL speech evaluation data released

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

[7] Appen refers to this set as dev in their documentation.

- IL Announcement (12:00[8] ET Aug 7)
    - Identity of IL announced
    - Decryption keys for **KB**, **Set 0** and **Set E** released
- Evaluation Checkpoint 1 (12:15 ET Aug 7 – 12:00 ET Aug 10)
    - Train with data from **Set 0** begins
    - Submission due at the end of Evaluation Checkpoint 1
    - At the end of Evaluation Checkpoint 1 and after submission is made[9], decryption keys for **Set 1** and **Set S** released
- Evaluation Checkpoint 2 (12:15 ET Aug 7 – 12:00 ET Aug 17)
    - Train with data from **Set 0** begins
    - Train with data from **Set 1** and **Set S** begins
    - Submission due at the end of Evaluation Checkpoint 2
    - At the end of Evaluation Checkpoint 2 and after submission is made, decryption keys for **Set 2, Set 0 Speech,** and **Set E Speech** released
- Evaluation Checkpoint 3 (12:15 ET Aug 7 – 12:00 ET Aug 24 for MT, EDL, SF Text and 12:00 ET Aug 31 for SF Speech)
    - Train with data from **Set 0** begins
    - Train with data from **Set 1** and **Set S** begins
    - Train with data from **Set 2** (and **Set 0 Speech** if applicable) begins
    - Submission due at the end of Evaluation Checkpoint 3

## 6.2 Data Description

The composition of the KB and datasets (**KB, Set 0, Set 1, Set 2, Set S, Set E, Set 0 Speech, Set E Speech**) for each incident language are listed in Table 1 below. 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 above (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 dataset described above 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.

[8] Military time format 00:00 - 24:00; 12:00 means noon.

[9] Valid and scorable submission at the current checkpoint and checkpoint deadline open the next checkpoint. Therefore, if an open participant chooses not to participate in a checkpoint, he/she must contact NIST to open the checkpoint for him/her. NIST will also open the checkpoint for SF Speech participants.

**Table 1: LoReHLT17 IL data description**

| **Set 0 – pre-incident epoch** |
|---|
| Category I Resources[10]<br>● Monolingual Source Text:<br>o ~100Kw newswire<br>o ~75Kw discussion forum/blog<br>o ~50Kw Twitter/SMS<br>● Parallel Text[11]:<br>o ~100Kw newswire<br>o ~100Kw discussion forum/blog<br>o ~100Kw Twitter/SMS<br>● Parallel Dictionary (~10,000 stems/lemmas)<br>Category II Resources (any 5 of the following):<br>● parallel dictionary IL --> non-English<br>● monolingual IL dictionary<br>● monolingual IL grammar book<br>● parallel English --> IL grammar book<br>● monolingual IL primer book<br>● monolingual IL gazetteer<br>● parallel IL --> English gazetteer |
| **Set 1 – incident/post-incident epoch** |
| Monolingual Source Text – 1/3 of leftover after **Set E** is met |
| **Set 2 – incident/post-incident epoch** |
| Monolingual Source Text – 2/3 of leftover after **Set E** is met |
| **Set S – incident/post-incident epoch** |
| English Scenario Model – approximately 50Kw, genre balance will vary based on availability |
| **Set E – incident/post-incident epoch** |
| Source Text:<br>● ~100Kw newswire<br>● ~50Kw discussion forum/blog<br>● ~50Kw Twitter/SMS |
| **Set 0 Speech – incident/post-incident epoch** |
| There is no guarantee that incident(s) in speech will correspond to incident(s) in text. |

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

[11] 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.

| Set E Speech – incident/post-incident epoch |
| --- |
| There is no guarantee that incident(s) in speech will correspond to incident(s) in text. |

# 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[12] 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.

  MT, EDL, SF Text:
  - 1 hour for Evaluation Checkpoint 1
  - 5 hours for Evaluation Checkpoint 2 (4 hours if 1 hour was used in Checkpoint 1)
  - 5 hours for Evaluation Checkpoint 3

  SF Speech:
  - 10 hours for Evaluation Checkpoint 3

# 8 Evaluation Protocol

## 8.1 Evaluation Account

All evaluation activities will be conducted online via the evaluation account. Go to https://lorehlt.nist.gov to sign up for an account if you do not have one already. Participants will need a valid email address and choose a password that is at least 12 characters long including uppercase and lowercase letters, numbers, and special characters.

After signing up and confirming the account, each participant[13] will be asked to associate himself/herself

[12] 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.

[13] A *participant* is defined as a member of an organization who takes part in the evaluation (e.g., John Doe).

to a site[14] (or create his/her site if it does not exist). The first person who creates the site will be deemed the *site representative* and will have to approve participants who want to join his/her site. The site representative will be asked to associate his/her site to a team[15] (or create his/her team if it does not exist). The first person who creates the team will be deemed the *team representative* and will have to approve sites who want to join his/her team. The site representative can create other teams as well as ask to join his/her site to other teams. The team representative must register his/her team for a particular task to participate in that task. If the site declares itself as a LORELEI performer, its status will be verified. If the site is not a LORELEI performer, the site representative will be asked to sign the LDC license. The LDC will confirm the license and release the appropriate data to the site.

## 8.2 System Input File Format

With the addition of the Speech SF task, LoReHLT17 has two input source formats.

### 8.2.1 Input source for MT, EDL, SF Text

The input source data for the MT, EDL, and SF Text tasks follows a common data format LDC LTF 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 source for SF Speech

The input source data for the SF Speech 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.

[14] A *site* is defined to be a single organization participating in the evaluation (e.g., NIST).
[15] A *team* is defined to be a group of organizations collaborating on a task in the evaluation (e.g., NIST_LDC).

## 8.4 File List for MT, EDL, SF Text

The terms of usage of the Twitter data require that only the URLs of the tweets can be redistributed, not the actual tweets. Tweets can be deleted at any given time. **Participants are encouraged to harvest the tweets as soon as possible upon receipt of the evaluation data after the decryption keys are released.** As such, to distinguish between no output due to deleted tweets from no output due to a system's inability to produce the results, each team participating in MT, EDL, and SF Text is required to submit a file list along with their system output to indicate the source data availability. Even though this issue only affects the Twitter data, we ask teams to submit a list indicating the availability of all files in **Set E** for ease of use. For consistency, use the file list distributed with Set E (called '`filelist.txt`') and add a new field to indicate the file availability.

```
<DocID><tab><Available>
```

For example:

```
DF_AOA_TUR_0000116_20140900 TRUE
SN_TWT_TUR_2221137_20141021-02   FALSE
```

## 8.5 Submission Requirements

All teams are required to participate in the constrained training condition and are encouraged to participate in the unconstrained training condition as well. Submissions will not be classified as primary or contrastive in LoReHLT17.

One of the goals of the LoReHLT evaluation is to track system performance over time. As such LORELEI performers are required to submit at least one complete ensemble under the constrained training condition for each IL. An *ensemble* is defined to be a set of three submissions, one at each checkpoint, that are deemed comparable over time. Open participants are not required to submit a complete ensemble. As with LORELEI performers, they can only submit their system output for a particular checkpoint while it is open.

Teams may upload up to a maximum of 10 submissions per checkpoint under each training condition for each IL participated. All submission slots will be preassigned along with predefined ensemble labels in the format of IL{5,6}-(un)constrained-{1...10}. Unlike in LoReHLT16, there is no need to fill up available submission slots and so it is not recommended to submit the same system output more than once at any given checkpoint since all valid submissions will be reported and teams will be given an opportunity to rearrange the submissions across the checkpoints after checkpoint 3 closes.

By default, IL5-constrained-1 and IL6-constrained-1 will be considered as the aforementioned required ensembles for LORELEI performers. If for whatever reason a team prefers different designations but is unable to change them when managing submission assignments, please contact NIST.

Results for 10% of the evaluated portion will be given at submission time. Teams may use the results on the 10% to inform their future submissions rather than to replace an existing submission. 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. Otherwise, the new submission will count toward the 10 submission limit. Please note that while the 10% is planned to be proportionate to the full evaluated portion in terms of domain and genre distributions, it is **not guaranteed** to match proportionately to the number of SFs, entities, etc. of the full evaluated portion since the full annotation will not be completed by the time the selection of the 10% is to be made.

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. The system descriptions will be compiled into the workshop proceedings. Teams can download the template for the system description on the NIST LoReHLT17 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[16].
- 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 three 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 interchange with the outside world. 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.

[16] 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' ability to handle a new language. 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).
- Dry run of EDL includes only format validation (no scores)
- No dry run for SF Speech

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 & Ablation Study (LORELEI Performers Only)

LORELEI performers are required to reprocess the LoReHLT16 evaluation test set for the two tasks (MT and NER[17]). 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 4 *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 hours) teams use their prepared components to process the evaluation set. Teams can also use data gathered from the extra 1-hour 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 evaluation in August.
- 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 hours to process the test data and submit the results. Performers may make as many submissions as they wish. There is no checkpoint and no feedback of results.
- LORELEI performers will be provided some time with a native informant. Each team will have up to one hour 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.

Immediately following the retest, there will be an optional ablation study. Teams are to submit a single primary submission during the 24-hour retest and are encouraged (but not required) to participate in an ablation study that follows the retest. Teams will have 5 days to submit additional runs for the ablation study. It is up to the teams to define the parameters of their ablation experiment. DARPA is most interested to learn about language independent techniques, language projection techniques, and novel uses of native informant.

# 13 LoReHLT Schedule (tentative)

| ***Milestone*** | ***Date*** |
|---|---|
| Initial version of evaluation plan published | Dec 12, 2016 |
| Registration period | Mar 1 – May 31, 2017 |
| 6-month PI meeting (LORELEI performers only) | TBD |
| Uyghur retest (see ***Uyghur Retest Schedule*** below) | Jul 2017 |
| Dry run evaluation (see ***Dry Run Schedule*** below) | Jul 2017 |
| Official evaluation period (see ***Official Evaluation Schedule*** below) | Aug 2017 |
| DARPA PI meeting (LORELEI performers only) | Sep 12 – 14, 2017 |

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

| | |
|---|---|
| NIST post-evaluation workshop co-located with TAC/TREC | TBD |
| ***Uyghur Retest & Ablation Study Schedule (LORELEI Performers Only)*** | |
| Evaluation data available[18] | 12:00 ET Jul 11 |
| System output submission for retest due | 12:00 ET Jul 12 |
| System output submission for Ablation Study opens | 12:15 ET Jul 12 |
| System output submission for Ablation Study closes | 12:00 ET Jul 17 |
| ***Dry Run Schedule*** | |
| Encrypted data released by LDC | Jul 17 |
| IL Announcement<br>- Decryption keys for set 0 and set E distributed<br>- System description submission opens<br>- Submission for checkpoint 1 opens | 12:00 ET Jul 18 |
| Evaluation Checkpoint 1<br>- System description submission opens<br>- System output submission for Evaluation Checkpoint 1 opens<br>- Decryption key for set 1 and set S distributed at end of Evaluation Checkpoint 1 and after system output submission made | 12:15 ET Jul 18<br>12:00 ET Jul 19 |
| Evaluation Checkpoint 2<br>- System output submission for Evaluation Checkpoint 2 opens<br>- Decryption key for set 2 distributed at end of Evaluation Checkpoint 2 and after system output submission made | 12:15 ET Jul 19<br>12:00 ET Jul 20 |
| Evaluation Checkpoint 3<br>- System output submission for Evaluation Checkpoint 3 opens | 12:15 ET Jul 20<br>12:00 ET Jul 21 |
| System description submission closes | 12:15 ET Jul 21 |
| Preliminary results released if system description is received | Jul 24 |
| ***Official Evaluation Schedule*** | |
| Encrypted data released by LDC | Aug 04 |
| IL Announcement<br>- Decryption keys for set 0 and set E distributed | 12:00 ET Aug 07 |
| Evaluation Checkpoint 1<br>- System description submission opens<br>- Access to Native Informant (MT, EDL, SF Text; see below)<br>- System output submission for Evaluation Checkpoint 1 opens<br>- Decryption key for set 1 and set S distributed at end of Evaluation Checkpoint 1 and after system output submission made | 12:15 ET Aug 07 12:00 ET Aug 10 |
| Evaluation Checkpoint 2<br>- Access to Native Informant (MT, EDL, SF Text; see below)<br>- System output submission for Evaluation Checkpoint 2 opens<br>- Decryption key for set 2 (and set 0 speech if applicable) distributed at end of Evaluation Checkpoint 2 and after system output submission made | 12:15 ET Aug 7<br>12:00 ET Aug 17 |
| Evaluation Checkpoint 3 | 12:15 ET Aug 7<br>12:00 ET Aug 24 (MT, EDL, SF Text) |

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

| | |
|---|---|
| - Access to Native Informant (MT, EDL, SF Text, SF Speech; see below)<br>- System output submission for Evaluation Checkpoint 3 opens | 12:00 ET Aug 31 (SF Speech) |
| System description submission closes[19] | 12:00 ET Aug 25 (MT, EDL, SF Text)<br>12:00 ET Sep 1 (SF Speech) |
| System description reviewed by NIST | Aug 29 (MT, EDL, SF Text)<br>Sep 5 (SF Speech) |
| Preliminary results released if system description is received | Sep 1 (MT, EDL, SF Text)<br>Sep 6 (SF Speech) |
| Native Informant Timeline (time amount is per incident language per team per task (MT, EDL, SF Text) | |
| Up to 1 hour between 12:15 ET Aug 07 to 12:00 ET Aug 10<br>Up to 5 hours between 12:15 ET Aug 10 to 12:00 ET Aug 17<br>(or 4 hours if 1 hour was used between Aug 07 and Aug 10)<br>Up to 5 hours between 12:15 ET Aug 17 to 12:00 ET Aug 24 | |
| Native Informant Timeline (time amount is per incident language per team per task (SF Speech) | |
| Up to 10 hours between 12:15 ET Aug 7 to 12:00 ET Aug 31 | |

# 14 Machine Translation (MT) Evaluation Specifications

## 14.1 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 test set must be translated, even though only a subset of it will be scored in the machine translation evaluation.

## 14.2 Performance Measurements

BLEU will be the primary metric. BLEU scores at the system, document, and segment level will be calculated at each checkpoint. Scoring will be done against two human reference translations. Guidelines for how the reference translations were created are available.[20] Scoring will be done preserving case. Other 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. For the 10% feedback set, the full detailed scores (down to the segment level) will be reported any time feedback is provided.

NIST will continue to investigate additional automatic approaches geared towards measurement of successful translation of content relevant to the LORELEI task.

[19] 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.

[20] LoReHLT Translation Guidelines" version 1.1. LORELEI performers can access the annotation guidelines through this link. LoReHLT participants can request this document directly from the LDC by sending a request to lorelei-poc@ldc.upenn.edu.

## 14.3 System Output Format

MT systems are required to output the translation conforming to the lorehlt-mt-v1.2.dtd[21]. 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 navgation, search</seg>
      …
    </doc>
  </tstset>
</mteval>
```

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.

## 14.4 System Submission Format

The MT system output files as described in 12.3 along with the file list as described in Section 8.4 named '`filelist.txt`' 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`'.

# 15 Situation Frame (SF) Text Evaluation Specifications

## 15.1 Task Definition

Given a text document in the incident language, an SF system is required to automatically identify the 0 or more situation frames covered in the document. Each system-generated SF consists of a situation type, place mention, and (for some types) status variables.

- Situation Type: A situation frame must be labeled as one of the pre-defined types in the LDC's "Annotation Guidelines for LORELEI Situation Frames"[22]. There are two kinds of situations: situations involving a 'need' (e.g., food supply, evacuation, etc.) or situations involving an 'issue' (e.g., civil unrest, terrorism, etc.). Regardless of the kind, the SF system will return a string for the situation type and a confidence score.
    - **SFType**: a text string indicating the type of situation. 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.

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

[22] "Annotation Guidelines for LORELEI Situation Frames" version 3.0. LORELEI performers can access the annotation guidelines through this link. LoReHLT participants can request this document directly from the LDC by sending a request to lorelei-poc@ldc.upenn.edu.

- **TypeConfidence**: a numeric confidence value indicating the strength of evidence supporting the identified situation type for the SF, ranging from 0 to 1, inclusive.

- Place Mention: A situation occurs at a physical place, either a location or region. The SF system will identify the named entity mention, in terms of the character extent and entity type, where the situation takes place if the document contains a named entity mention. In the event that the system is confident that no place mention should be associated with the frame, the system is expected to return an empty mention (i.e. "{}"). Reference SFs will be scored regardless of the 'Proxy' tag for place annotation.
  - **Begin**: Starting character offset of the mention within the source document
  - **End**: Ending character offset of the mention within the source document
  - **EntityType**: The entity type for the mention, either "GPE" or "LOC". (NOTE: EntityType will not be evaluated during the 2017 evaluation.)
- Status Variables: Status variables indicate relevant context describing the situation.
  - The 'issue' situation frames are optionally accompanied by following status variable:
    - **Issue:** Either "current" or "not_current"

    The value for this variable should be defaulted to "currently relevant", unless clearly specified.
  - The 'need' situation types are optionally accompanied by three status variables for each SF: "Need", "Relief", and "Urgency". The fill of each status variable is limited to an enumerated set prescribed by the annotation document. The system SF will list the following fills
    - **Need**: One of "current", "future", "past"
    - **Relief**: Either "sufficient" or "insufficient"
    - **Urgency**: Either true or false

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 **SFType** to be evaluated. Systems specifically not addressing the geographic localization and/or status variables will not be evaluated with respect to the omitted fields.

## 15.2 Performance Measurements

The conceptual use of SF technology is to support down-stream 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 and high false alarm application scenarios and high quality supporting information.

This year's SF evaluation will not address the aggregation test case directly. Rather, system performance will be measured by their ability to correctly identify the right number of SFs using SF equivalency classes to assess performance at several levels of granularity while using a single system output. The assessment procedure will also not require systems to perform within-document entity co-reference by not penalizing a system for generating multiple SFs that identify mentions of the same reference entity.

The equivalence classes are as follows:

- **SFType** (required): Only the **SFType** field is considered when comparing against the reference
- **SFType+Place**: Both **SFType**, and **Place** are considered
- **SFType+Place+Status**: **SFType**, **Place**, **Status** are considered
- **SFType+Place+Relief** (need frames only): **SFType**, **Place**, **Relief** are considered
- **SFType+Place+Urgency** (need frames only): **SFType**, **Place**, **Urgency** are considered

Following the results from the 2016 evaluation, we will be using multiple references for scoring for this year's evaluation. This change is due to the fact that the task seems to be more similar to machine translation, than retrieval, in that there are multiple plausible interpretations of the events (situations). We're also moving away from the SFError metric used in the 2016 evaluation, and instead will be using two metrics: F1, and what we call Occurrence Weighted F. The primary measure will be a range of F1 scores (an F1 score for each individual reference), while Occurrence Weighted F will be reported as a contrastive measure and uses an aggregated reference. Occurrence Weighted F is the harmonic mean of 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.

Occurrence Weighted F:

$$2 \cdot \frac{Occurrence\ Weighted\ Precision \cdot Occurrence\ Weighted\ Recall}{Occurrence\ Weighted\ Precision + Occurrence\ Weighted\ Recall}$$

Occurrence Weighted Precision:

$$\frac{\Sigma\, \alpha_{TP}}{\Sigma\, \alpha_{TP} + \Sigma\, \alpha_{FP}}$$

Occurrence Weighted Recall:

$$\frac{\Sigma\, \alpha_{TP}}{\Sigma\, \alpha_{TP} + \Sigma\, \alpha_{FN}}$$

Where $\Sigma\, \alpha_{TP}$ , $\Sigma\, \alpha_{FP}$ , and $\Sigma\, \alpha_{FN}$ are the sums of weights for the true positives, false positives, and false negatives respectively.

## 15.3 Scoring Procedure

For scoring, we first map the system and reference frames to "equivalence class frames", using each of the equivalence classes described above, then we compare the sets of system and reference frames to compute the metrics. More specifically:

1. For each system output situation frame, provided character extents (if place mention fields are included in that particular situation frame) are mapped to the reference mention, constituting **Place**. If no reference mention is found for the provided character extents, the **Place** is considered "Unknown". For the purposes of matching and reducing equivalence class frames, "Uknown != "Unknown".

2. For each **equivalence class**:
   2.1. Skip if the **equivalence class** is optional and the system produced no responses for the relevant fields
   2.2. Reduce the place-mapped system output situation frames (S) to equivalence class frames for the **equivalence class**, removing duplicates, giving us S′
   2.3. Reduce each set of reference situation frames (e.g. R1, R2, R3) to distinct sets of equivalence class frames R1′, R2′, and R3′ by applying the **equivalence class**, and removing duplicates
   2.4. Take the union of our individual reference equivalence class frames (R1′, R2′, and R3′), while also tracking the number of individual references containing each frame, giving us WR. $\{(\alpha, r) \in WR\}$
   2.5. Compute F1 using S′ and each of R1′, R2′ and R3′
   2.6. Compute Occurrence Weighted F using S′ and WR

## 15.4 System Output Format

The system output structure is a JSON structure and should conform to the json schema. The latest schema (LoReHLT17_v1.1), along with the latest LoReHLT Frame Scorer software package (v2.0.1), can be downloaded from the official LoReHLT '17 webpage. Contained below is a simple example of the system output structure.


```
[
  { "DocumentID": "CMN_NG_000031_20080707_80020000G",
    "Type": "infra",
    "TypeConfidence": 0.4,
    "PlaceMention": {
      "EntityType": "GPE",
      "Start": 28,
      "End": 29
    },
    "Status": {
      "Need": "current",
      "Relief": "insufficient",
      "Urgent": false
    }
  },
  { "DocumentID": "CMN_NG_000031_20080707_80020000G",
    "Type": "shelter",
    "TypeConfidence": 0.6,
    "PlaceMention": {
      "EntityType": "GPE",
      "Start": 212,
      "End": 213
    },
    "Status": {
      "Need": "current",
      "Relief": "insufficient",
      "Urgent": false
    }
```

```
    }
]
```

## 15.5 System Submission Format

The SF system output files as described in Section 15.4 named 'system_output.json' along with the file list as described in Section 8.4 named 'filelist.txt' should be placed into a 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) Speech Evaluation Specifications[23]

## 16.1 Task Definition

Given an audio segment in the incident language an SF system is expected to automatically identify any situation frames covered in the segment. A complete SF includes a document id, situation type, localization (optional) and a confidence score.

- Document ID: the file name of the corresponding audio segment (without extension)
- Situation Type: is a string corresponding to one of the pre-defined types, as defined in the Appen annotations.
- PlaceMention (optional): is a string – in the incident language script – indicating the physical place where the situation occurs.
- TypeConfidence: a number in [0,1] indicating the system's confidence that the frame exists. This is mandatory to allow for a curve-based evaluation.

Each system is expected to process all audio segments in a set and produce the corresponding frames.

## 16.2 Performance Measurement

In order to facilitate the creation of systems that can perform at various operating points, we will be performing a curve based evaluation. We will be using Precision-Recall (PR) curves, which allow the approach to generalize to the localization level (ROC and DET curves can not, due to the requirement for a True Negative estimate). For each system submission & for each layer of the evaluation a PR curve will be generated, with each point of the curve corresponding to a combination of micro-averaged recall and precision.

The curve will be produced by sweeping across the confidence values in the system output (using 500 percentiles at 0.2 intervals). Additionally, as an aggregate metric we will report the Area Under the Curve (AUC).

The process to estimate a single point on the PR curve is as follows:

1. Remove all frames below the current confidence threshold

[23] We would like to thank Shrikanth Narayanan and Nikolaos Malandrakis of the Signal Analysis and Interpretation Laboratory (SAIL) at the University of Southern California for running the 2016 pilot evaluation for Speech Situation Frame and for providing the evaluation plan material and scoring software for use in the 2017 Speech SF evaluation.

2. Transform the remaining frames to the current evaluation layer, by removing extraneous attributes and merging duplicates.
3. Align the ground truth and output frames via maximum similarity
4. Calculate True Positives, False Positives and False Negatives
5. Calculate Precision and Recall

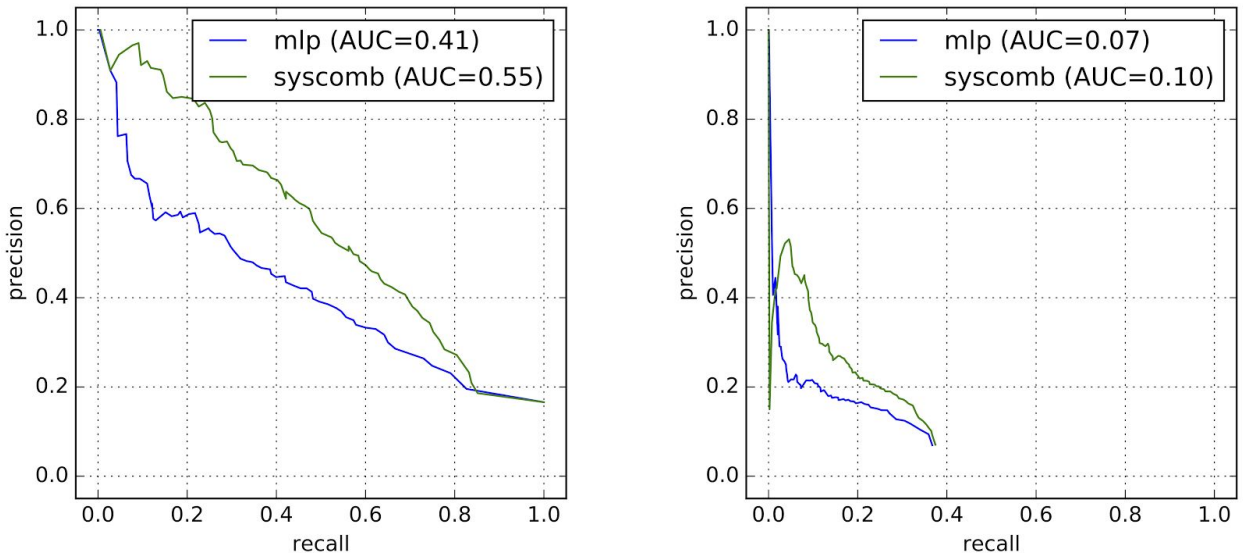



Figure 1: PR curve examples, for (a) Type and (b) Type+Place, for 2 systems

Figure 1 shows two examples of PR curves at the Type and Type+Place layers. Note that the Type+Place curve never reaches recall of 1.0; that is expected and part of why we will be conducting visual comparisons of these curves rather than depending solely on AUC.

To allow for the creation of these curves, we encourage the submission of low confidence results. For "Type", participants are advised to produce all possible Types for every segment, even if they have a confidence score of zero.

### 16.2.1 Evaluation Layers

For the purposes of this evaluation we consider the following layers.

1. Relevance: Does this segment contain at least 1 frame of any type? For this class all attributes are discarded, except for the document id.
2. Type: Which (if any) types of frames are contained in the segment? For a frame to be correct at this layer, it has to have the correct document id and type.
3. Type+Place: Which (if any) types of frames are contained in the segment and where are they localized? For a frame to be correct at this layer it needs to have the correct document id, type and location. Note that non-localized frames are ignored at at this layer.

Each participant will only need to submit a single output to be evaluated on one or more of these layers in order.

- An output containing localized frames will be evaluated on all 3 layers.
- An output not containing any localized frames, but including actual Types will be evaluated for Type and Relevance.

### 16.2.2 Frame Similarity

To allow for partial credit at the localization level, we are introducing the concept of frame similarity, indicated by a number in [0,1] with 1 indicating a perfect match.

For the Relevance and Type layers of the evaluation the calculation is trivial: the frames are either perfectly matched or not, giving the similarity metric values of 1 and 0 respectively. For the Type+Place layer, we will be using a soft matching of the PlaceMention strings and the similarity between two frames (if Type and Document ID match) will be equal to that string similarity measure.

String similarity is defined as the character-level edit distance between the two PlaceMentions, normalized by the sum of their string lengths:

Similarity = (sum(length) - minimum edit distance)/sum(length)

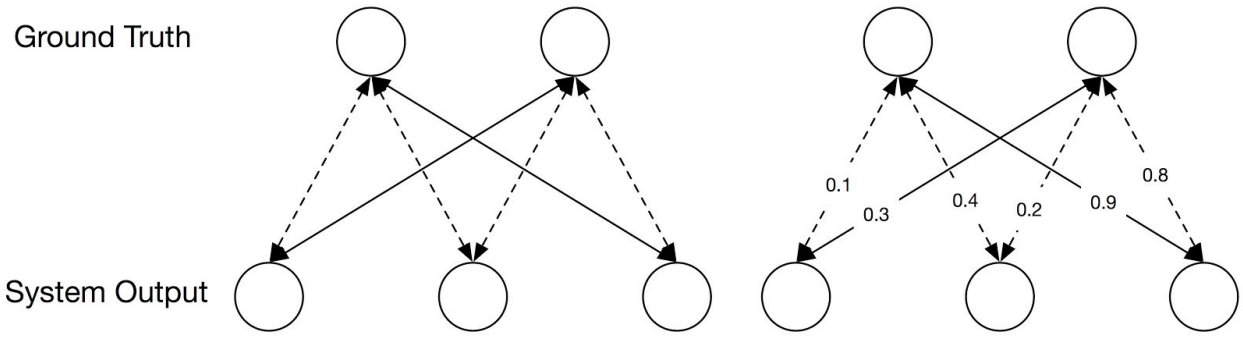
This metric takes values in [0,1]. The edit distance is calculated using costs of 1 for insertions and deletions and 2 for substitutions.

### 16.2.3 Frame Alignment

The frames in the ground truth and system output are aligned using a maximum similarity criterion. All pair-wise similarities are calculated and, using a linear assignment algorithm, each frame in the output is mapped to 0 or 1 frames in the ground truth in such a way as to maximize the sum of similarities.

In the mappings, no frame may be matched more than once.

An example is shown below, for the case of hard and soft matching.




The solid arrows represent the frame alignment and, in the case of soft matching, the arrows have similarity scores on them.

The scoring takes into account the similarity scores and gives partial credit, by using soft set cardinality.

For the hard matching example, the scoring would be:

- True positive = 2
- False negative = 0
- False positive = 1

Whereas the soft matching example would yield:

- True positive = 0.9+0.3 = 1.2
- False negative = 2 (reference cardinality) - 1.2 = 0.8
- False positive = 3 (output cardinality) - 1.2 = 1.8

## 16.3 Output Format

The system output is a single json file with a structure that adheres to the schema (the schema file can be found on the official LoReHLT '17 evaluation web page). Note that while the schema allows for the inclusion of the status variables "Need" and "Relief", they will not be evaluated during the first year pilot of the task.

A complete frame would look like this:


```
{
"DocumentID": "CHN_EVAL_096_004",
"PlaceMention": "\u6c5f\u82cf",
"Type": "Medical Assistance",
"TypeConfidence": 0.5585732473158215
}
```


Note the unicode encoding of the "PlaceMention" string. A valid system output can use either proper Unicode characters in the native script or their u-code versions.

### 16.3.1 Frame examples - with layers

A complete frame, including status variables (which will be ignored during the evaluation).


```
{
"DocumentID": "CHN_EVAL_096_004",
"PlaceMention": "\u6c5f\u82cf", "Status": {
      "Need": "Past Only",
      "Relief": "No_Known_Resolution"
},
"Type": "Medical Assistance",
"TypeConfidence": 0.5585732473158215
}
```


A localized frame.


```
{
"DocumentID": "CHN_EVAL_096_004",
"PlaceMention": "\u6c5f\u82cf",
"Type": "Medical Assistance",
```

```
"TypeConfidence": 0.5585732473158215
}
```

A non-localized frame. This is the minimum information required for a frame to be valid.

```
{
"DocumentID": "CHN_EVAL_096_004",
"Type": "Medical Assistance",
"TypeConfidence": 0.5585732473158215
}
```

## 16.4 The Appen Annotations and Special Cases

The Appen annotations look like this:

```
TYPE: Type1
TIME: Past Only
Resolution: Sufficient
PLACE: Place1
```

Each annotation includes these 4 lines and each audio segment may correspond to multiple of these 4 line combinations. However, these lines may include multiple Types and locations. For example:

```
TYPE: Type1, Type2
TIME: Past Only
Resolution: Sufficient
PLACE: Place1
```

This, for the purposes of this evaluation, counts as two frames, both localized to Place1, with Types being Type1 and Type2. In the cases where there is one Type and multiple locations or multiple Types and zero or one locations we consider each possible combination of Type and location as a separate frame.

A special case is when this structure contains multiple Types and multiple locations, like below:

```
TYPE: Type1, Type2
TIME: Past Only
Resolution: Sufficient
PLACE: Place1, Place2
```

This is meant be read as: “Type1 at Place1 or Place2 or both” and “Type2 at Place1 or Place2 or both”. So each type may be connected to either or both types, it is ambiguous.

It is clear how to evaluate this at the “Type” layer: all types must be assigned to the segment. It is not clear how we may evaluate at the “Type+Place” layer, due to the ambiguity: if a system output contains “Type1 at Place2”, we do not know if that is correct, since Type1 may only apply to Place1. Only a very small percentage of all annotations fall under this special case, so we ignore these segments when evaluating at the Type+Place layer.

However, these segments will be taken into account when evaluating at the Type and Relevance layers.

## 16.5 System Submission Format

The JSON formatted system output file as described in subsection 16.3, should be named 'system_output.json', and placed in a flat-file hierarchy and compressed into either a .tgz or .zip file. 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), an EDL system is required to automatically identify and classify entity mentions into pre-defined 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.

As with the NER task in LOREHLT16, in the LOREHLT17 EDL task, the mention type is limited to named mentions only and the entity types are limited to 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 NER, 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

### 17.1.1 Knowledge Base (KB)

The reference KB – all in English – 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 has been distributed for participants to become familiar with the format, which includes a few examples of manually augmented entries, unrelated to any IL's to avoid exposing evaluation-sensitive information.

## 17.2 Performance Measurements

Scoring metrics from the TAC KBP2015/2016 EDL task will be extended to the EDL task. Specifically, Precision, Recall and F1 scores will be reported for the following metrics (applicable to both 10% feedback and final scores):

**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
- typed_mention_ceaf
- typed_mention_ceaf_plus

**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.

Please note that for the 10% feedback score, the reference data will not include cross-document NIL coreference.

## 17.3 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 TAC2015/2016 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 or NIL clustering ID

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

Field 7: all should be of type {NAM}

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.

## 17.4 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).