NIST LoReHLT 2017 Evaluation Plan

Last Updated March 6, 2017

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 aspects, 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 datase; 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.
- **Text Situation Frame** – for each document, automatically generate Situation Frames covered in the document. For text SF specific requirements, see Section 15.
- **Speech Situation Frame**[2] – for each audio recording, automatically generate Situation Frames covered in the recording. For speech SF specific requirements, see Section 16.

[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.

- **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.

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 enable 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 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 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 use during the evaluation so long as they are multilingual-focused in nature. Teams will be also given some resources to use described in Section 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 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 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 6 is permitted. Prior to the incident and incident language announcement, teams can assemble 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**.

[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.

- **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 6[4]. 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 **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 at lorelei-poc@ldc.upenn.edu and ask for the LoReHLT translation, situation frame, or entity discovery and linking 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

All three evaluation tasks will use the same data component and have the same release plan. The LDC releases the Incident Language (IL) data and English Scenario Model in an encrypted format (see Section 6.4), and NIST releases the appropriate decryption key(s) at the appropriate stages for each IL. The stages are:

- Pre-IL Announcement (before the IL Announcement)
    - **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 evaluation data released
- IL Announcement
    - Identity of IL announced
    - Decryption keys for **set 0** and **set E** released
- Evaluation Checkpoint 1
    - Train with data from **set 0** begins at IL Announcement
    - Evaluation Checkpoint 1 submission due 3 days after IL Announcement

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

    - Decryption key for **set 1** and **set S** released 3 days after IL Announcement and after submission to Evaluation Checkpoint 1 made[5]
- Evaluation Checkpoint 2
    - Train with data from **set 0** begins at IL Announcement
    - Train with data from **set 1** and **set S** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
    - Evaluation Checkpoint 2 submission due 10 days after IL Announcement
    - Decryption key for **set 2** released 10 days after IL Announcement and after submission to Evaluation Checkpoint 2 made
- Evaluation Checkpoint 3
    - Train with data from **set 0** begins at IL Announcement
    - Train with data from **set 1** and **set S** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
    - Train with data from **set 2** begins after the Evaluation Checkpoint 2 submission deadline and the team makes a submission
    - Evaluation Checkpoint 3 submission due 17 days after IL Announcement

6.2 Data Description

The composition of the five datasets (**set 0, set 1, set 2, set S, and set E**) 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 with another genre. “Kw” refers to multiples of 1000 words.

6.3 Data Format and Structure

These five datasets (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.

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[6] 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.

[5] 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.

[6] 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.

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.
- 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 hour for Evaluation Checkpoint 1
    - 5 hours for Evaluation Checkpoint 2 (4 hours if 1 hour was used in Checkpoint 1)
- Teams cannot ask the native informant about the incident regardless of the training condition.

Table 1: LoReHLT16 IL data description

Set 0 – pre-incident epoch

Category I Resources[7]

- Monolingual Source Text:
  - ~100Kw newswire
  - ~75Kw discussion forum/blog
  - ~50Kw Twitter/SMS
- Parallel Text[8]:
  - ~100Kw newswire
  - ~100Kw discussion forum/blog
  - ~100Kw 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 – 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:

- ~100Kw newswire
- ~50Kw discussion forum/blog
- ~50Kw Twitter/SMS

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

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

[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.

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[9] will be asked to associate himself/herself to a site[10] (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[11] (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

The input source data to the system is the same across all three tasks and uses the LDC LTF format conforming to the LTF DTD referenced inside the test files. For a detailed description of the evaluation IL package, see Table 1.

Each team is to process the entire test set even though for some tasks only a portion of the test will be scored. 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>
```

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

[10] A *site* is defined to be a single organization participating in the evaluation (e.g., NIST).

[11] A *team* is defined to be a group of organizations collaborating on a task in the evaluation (e.g., NIST_LDC).

8.3 System Output File Format

While all tasks have the same system input file format, each has its own output format. Refer to the task specific section for information about the output requirement for that task.

8.4 File List

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 is required to submit a file list along with their system output to indicate the source data availability. Even though this issue is only affected 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

Teams are required to participate in the constrained training condition and are encouraged to participate in the unconstrained training condition. 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 ensemble per the training condition participated. An *ensemble* is defined to be a set of three submissions, one at each checkpoint, that are deemed comparable over time. Open participants can participate at any or all checkpoints.

Teams have a maximum of 10 submissions per checkpoint. 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 who will enable the 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 distribution, 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.

Submissions will not be classified as primary or contrastive in LoReHLT17. All valid submissions per checkpoint will be reported. At each submission, teams are required to provide a short description of their submissions after they upload their system output. At the conclusion of the evaluation, teams will be asked to connect the submissions across the checkpoints to form ensembles if applicable.

As stated previously, LORELEI performers are required to have at least one complete ensemble. All teams are required to submit a more formal system description that covers their submissions for all tasks the team participated 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[12].
- 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. Failure to attend the workshop may result in participant being denied from participating in future evaluations.
- 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.

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.

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

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 MIG 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, as 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 as much as possible. Some of the differences between the dry run and the official evaluation are:

- Shorter time duration between checkpoints
- No native informant
- The identity of the language is known before the IL Announcement (Mandarin same data as last year).
- Dry run of EDL includes only format validation (no scores)

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 are required to reprocess the LoReHLT16 evaluation test set with their 2017 systems for the two tasks (MT and NER[13]). The retest is an informal assessment[14] of progress made on cross-language techniques. 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 the native informant. Each team will have up to one hour with a native informant per task. No additional time with the native informant is allowed before or during the retest, even at the performers' cost.

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 below)	Jul 2017
Dry run evaluation (see below)	Jul 2017
Official evaluation period (see below)	Aug 2017
DARPA PI meeting (LORELEI performers only)	TBD
NIST post-evaluation workshop co-located with TAC/TREC	TBD
Uyghur Retest Milestone	
Evaluation data available[15]	Noon ET Jul 11
Submission due	Noon ET Jul 12
Dry Run Schedule	

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

[14] Informal because the conditions of the retest are not the same, e.g., the identity of test language is not truly unknown but performers are to act as if it were.

[15] LORELEI performers should have the evaluation data already. Open participants will need to download the evaluation data from the LDC.

Encrypted data released by LDC	Jul 17
IL Announcement - Decryption keys for set 0 and set E distributed by NIST - System description submission opens - Access to Native Informant begins - Submission for checkpoint 1 opens	Noon ET Jul 18
Evaluation Checkpoint 1 - Access to Native Informant ends - Submission for checkpoint 1 closes - Decryption key for set 1 and set S distributed after submission made - Submission for checkpoint 2 opens	Noon ET Jul 19
Evaluation Checkpoint 2 - Submission for checkpoint 2 closes - Decryption key for set 2 distributed after submission made - Submission for checkpoint 3 opens	Noon ET Jul 20
Evaluation Checkpoint 3 - Submission for checkpoint 3 closes	Noon ET Jul 21
System description submission closes	Noon 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 - Decryption keys for set 0 and set E distributed by NIST - System description submission opens - Access to Native Informant begins - Submission for checkpoint 1 opens	Noon ET Aug 07
Evaluation Checkpoint 1 - Submission for checkpoint 1 closes - Decryption key for set 1 and set S distributed after submission made - Submission for checkpoint 2 opens	Noon ET Aug 10
Evaluation Checkpoint 2 - Access to Native Informant ends - Submission for checkpoint 2 closes - Decryption key for set 2 distributed after submission made - Submission for checkpoint 3 opens	Noon ET Aug 17
Evaluation Checkpoint 3 - Submission for checkpoint 3 closes	Noon ET Aug 24
System description submission closes	Noon ET Aug 25
System description reviewed by NIST	Aug 29
Preliminary results released if system description is received	Aug 31
Native Informant Timeline (time amount is per incident language per team per task)	
Up to 1 hour between noon ET Aug 07 to noon ET Aug 10 Up to 5 hours between noon ET Aug 10 to noon ET Aug 17 (or 4 hours if 1 hour was used between Aug 07 and Aug 10)	

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 metrics. BLEU scores will be calculated at each checkpoint. Scoring will be done against two human reference translations. 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.

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

14.3 System Output Format

MT systems are required to output the translation conforming to the lorehlt-mt-v1.2.dtd[16]. 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 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`'.

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

15 Situation Frame (SF) 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 localization, 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”[17]. There are two general classes 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 general class, the SF system will return a string for the situation type and a confidence score.
    - **SFType**: a text string indicating the enumerated type of situation.
    - **TypeConfidence**: a numeric confidence value indicating the strength of evidence supporting the identified situation type for the SF. (NOTE: TypeConfidence will not be evaluated during the 2017 evaluation.)
- 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 there is no named mention in the document, the system is expected to not return a mention. 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 types are not accompanied by status variables.
    - The ‘need’ situation types are 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 only”, “Past only”
        - **Relief**: One of “Sufficient”, “Insufficient/Unknown sufficiency”
        - **Urgency**: true | 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 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.

[17] “Annotation Guidelines for LORELEI Situation Frames”

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 initial 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.

In order to evaluate system performance, the following procedure will be performed for each document, for each entity type:

- Define the **equivalency class(es)** for the given metric:
    - The classes will describe which SF components to collapse in order to reduce the set of system frames. For example:
    - /place=place, need=*, relief=*, urgency=*/ treats SFs with differing status variables as equivalent.
    - /place=*, need=*, relief=*, urgency=*/ treats SFs with differing place and status variables as equivalent.
- Build the reduced set of scorable reference SFs (*R'*) using the equivalency classes and removing SFs with a 'true' proxy tag.
- Build the reduced set of scorable system SFs (*S'*) using the equivalency classes.
- Tally:
    - *Cor* = Correct SFs, the set of elements in R' with at least one matching S' based on the equivalency classes. Note: the definition of 'correct' is described below for each measure.
    - *Spu* = Spurious SFs, the set of elements in S' not matching any R' elements
    - *Del* = Deleted SFs, the set of elements in R' with no matching S' elements

The following metrics will be computed for the SFType, Place Mention, and Status Variables.

15.2.1 SFType Performance Measure

SFType performance will be measured as a 'recognition' task using Situation Frame Error (SFE) rate. The measure will answer: "Did the system produce the right 'type' of SFs for the document?" SFE is the ratio of spurious and deleted SFs to the number of reference SFs pooled over the test collection. For SFType performance, place mention and status variables for both system and reference SFs will be treated as equivalent.

Equivalence classes: /place=*, need=*, relief=*, urgency=*/

Correct SF requirements: The SFType of both system and reference SFs must match.

$SFE_{SFType} = |Spu + Del|/|R'|$

In addition to *SFE* and using the above equivalence class and correct SF requirement, we calculate *Precision* and *Recall* between system and reference SF annotations:

$Precision = | S' \cap R'|/|S'|$

$Recall = | S' \cap R'|/|R'|$

F1 will also be calculated as Precision and Recall harmonic mean:

F1 = 2 * Precision * Recall/(Precision + Recall).

SFE, Precision, Recall, and F1 will be calculated and reported over the full test collection, genre, SFType(s), Need SFType(s), and Issue SFTypes(s).

15.2.2 SFType+Place Mention Performance Measure

Joint SFType and Place Mention performance will be measured as Situation Frame Error (SFE) rate. The measure will answer: “Did the system produce the right set of ‘type+place’ SFs for the document?”. A system will not be penalized by creating multiple SFs for the same reference entity so long as the types match and the system’s place mention extent matches at least one mention extent of the reference entity’s mentions effectively ‘no-scoring’ the duplicates. For this measure, all status variables are treated as equivalent.

Equivalence classes: /place=place, need=*, relief=*, urgency=*/

Correct SF requirements: The SFType of both system and reference SFs must match and the system mention extent must match at least one mention of the reference entity’s mentions.

$SFE_{SFType+Place} = |Spu + Del|/|R'|$

Also, using the same equivalence class and correct SF requirement, Precision and Recall and F1 will be computed between system and reference SF annotations:

$Precision = | S' \cap R'|/|S'|$

$Recall = | S' \cap R'|/|R'|$

F1 = 2 * Precision * Recall/(Precision + Recall).

15.2.3 SFType+Place+Status Performance Measure

Joint SFType, Place Mention, and Status performance will be measured as Situation Frame Error (SFE) rate. The measure will answer: “Did the system produce the right set of ‘type+place+status variable X’ SFs for the document?” Each status variable will be evaluated separately (even though ‘need’ and 'urgency' are inter-related) using a separate equivalence class for each variable and applying the same place mention matching rules as above.

Type+Place+Need:

Equivalence classes: /place=place, need=need, relief=*, urgency=*/

Correct SF requirements: The SFType, place mention (as described in 13.2.2), and Need status of both system and reference SFs must match

$SFE_{SFType+Place+Need} = |Spu + Del|/|R'|$

Type+Place+Relief:

Equivalence classes: /place=place, need=*, relief=relief, urgency=*/

Correct SF requirements: The SFType, place mention (as described in 13.2.2), and Relief status of both system and reference SFs must match

$SFE_{SFType+Place+Relief} = |Spu + Del|/|R'|$

Type+Place+Urgency:

Equivalence classes: /place=place, need=*, relief=*, urgency=urgency/

Correct SF requirements: The SFType, place mention (as in 13.2.2), and Urgency status of both system and reference SFs must match

$SFE_{SFType+Place+Urgency} = |Spu + Del|/|R'|$

Also, for all above equivalence classes and correct SF requirements, Precision and Recall and F1 will be computed between system and reference SF annotations:

$\text{Precision} = |S' \cap R'|/|S'|$

$\text{Recall} = |S' \cap R'|/|R'|$

$F1 = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall})$.

In addition to all the discussed metrics above, other evaluation metrics may be added based on the outcome of SF annotation exercise.

15.3 System Output Format

The system output structure is a JSON structure and should confirm to the json schema “lorehlt-sf_output-schema_v0.2.json” that is available online[18]. Contained below is an initial example that is also available online[19].

```
[
  { "DocumentID": "123",
    "Type": "Water Supply",
    "TypeConfidence": 0.5,
    "PlaceMention": {
      "EntityType": "GPE",
      "Start": 25,
      "End": 40
    },
    "Status": {
      "Need": "Current",
      "Relief": "No known resolution",
      "Urgent": true
    }
  },
  { "DocumentID": "123",
    "Type": "Civil Unrest or Wide-spread Crime",
    "TypeConfidence": 0.7,
    "PlaceMention": {
      "EntityType": "LOC",
      "Start": 12,
      "End": 23
    }
  }
]
```

15.4 System Submission Format

The SF system output files as described in 13.3 named ‘system_output.json’ 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`’.

16 Speech Situation Frame (SF) Evaluation Specifications

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

[18] ftp://jaguar.ncsl.nist.gov/lorehlt16/lorehlt-sf_output-schema_v0.2.json

[19] ftp://jaguar.ncsl.nist.gov/lorehlt16/lorehlt-sf_sample-system-output_v0.2.json

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
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

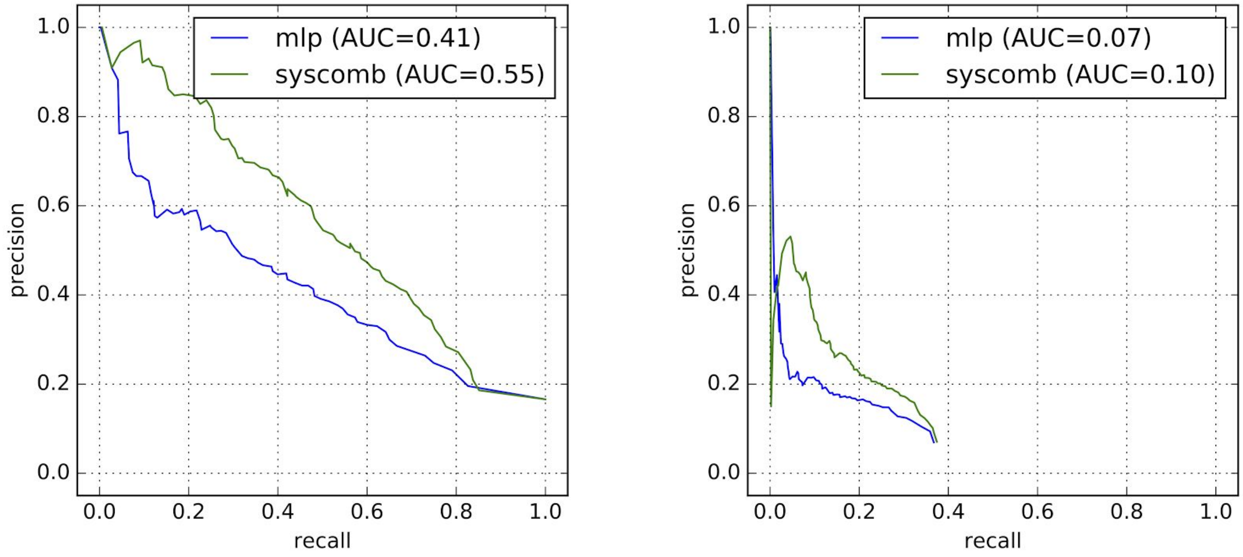


Figure 1: PR curve examples, for (a) Type and (b) Type+Place, for 2 systems

Figure 1 shows two examples of PR curves at the Type and Type+Place layers. Note that the Type+Place curve never reaches 1 recall; 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)

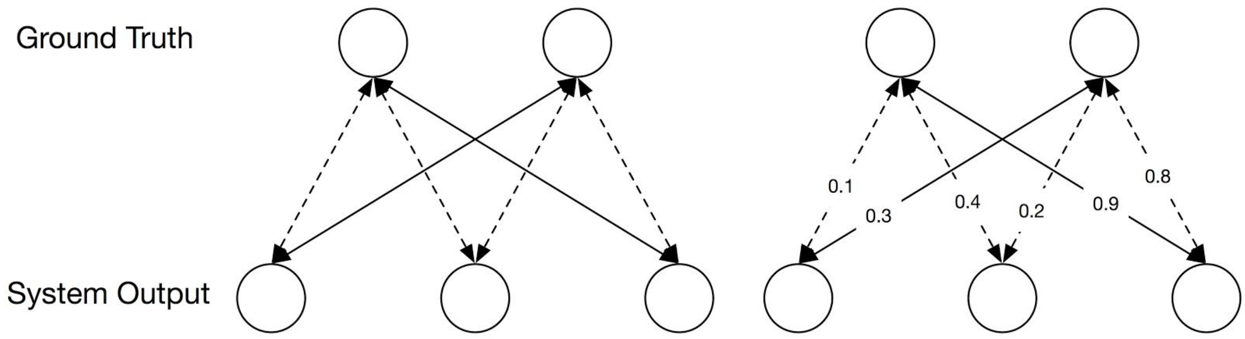
This metric takes values in [0,1]. The edit distance is calculated using costs of 1 for insertions and deletions and 2 for substitutions.

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.

The mappings are 1-to-1, 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 in the following 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.

The complete system output contains a list of Situation Frames, separated by commas and enclosed in square brackets (also see the attached evaluation script & sample output.

The JSON schema.

```
{
      "$schema": "http://json-schema.org/draft-04/schema#",
      "$version": "1.0",
      "definitions": {
            "frame": {
                  "type": "object",
                  "properties": {
                        "DocumentID": { "type": "string" },
                        "Type": { "type": "string",
                              "enum": [ "Civil Unrest or Wide-spread Crime",
                              "Elections and Politics",
                              "Evacuation",
                              "Food Supply",
                              "Infrastructure",
                              "Medical Assistance",
                              "Shelter",
                              "Terrorism or other Extreme Violence",
                              "Urgent Rescue",
                              "Utilities, Energy, or Sanitation",
                              "Water Supply" ] },
                "TypeConfidence": {  "type":  "number",  "minimum":  0,
            "maximum": 1 },
                "PlaceMention": { "type": "string" },
                "Status": {
                    "type": "object",
                    "properties": {
                        "Need": {
                            "type": "string",
                            "enum": [ "Current",
                            "Future",
                            "Past Only" ] },
                        "Relief": {
                            "type": "string",
                            "enum": [ "Insufficient/Unknown",
                            "No_Known_Resolution",
                            "Sufficient" ] }
                        },
                    "required": [ "Need", "Relief" ]
                }
            },
            "required": ["DocumentID", "Type", "TypeConfidence"]
        }
     },
    "type": "array",
    "items": {
        "$ref": "#/definitions/frame"
    }
}
```

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 1 Type & multiple locations or multiple Types & 0 or 1 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 our current plan (unless there is a better suggestion) is to ignore these segments when evaluating at the Type+Place layer.

They will be taken into account when evaluating at the Type and Relevance layers.

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).

LDC may not release EDL annotation guidelines specifically tailored for LOREHLT any time soon. Participants should 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

For more details on NER, please consult LDC's Simple Named Entity Annotation Guidelines. If you are an open participant and do not have direct access to the annotation guidelines, please contact LDC at lorelei-poc@ldc.upenn.edu.

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 may be distributed to prior to the start of the evaluation, but care will be taken not to reveal the identity of the ILs or any other evaluation-sensitive information.

17.2 Performance Measurements

Scoring metrics from the TAC KBP2015/2016 EDL task will be extended to the EDL task. A detailed description can be found in section 2.2 in the 2015 KBP overview paper at http://nlp.cs.rpi.edu/paper/kbp2016.pdf. The scorer is posted at https://github.com/wikilinks/neleval.

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
```

17.4 System Submission Format

Each afore-mentioned 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).