NIST LoReHLT 2016 Evaluation Plan

Last Updated February 11, 2016

1 Introduction

The DARPA Low Resource Languages for Emergent Incidents (LORELEI) Program seeks 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.

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 three evaluation tasks. LORELEI performers are required to participate in the tasks as outlined by their Statement of Work. Volunteer participants 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 13.
- **Topic Labeling (TL)** – the TL task specifics are still being discussed, but will be defined in Section 14.
- **Name Entity Recognition (NER)**[2] – for each document, identify and classify named mentions of PER, GPE, ORG, LOC entities. For NER specific requirements, see Section 15.

3 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 technology. The intent of the 'constrained' training condition is to test multilingual systems that are re-targeted to a incident language using a fixed amount of incident language materials. Teams should consult with NIST if their approach is not easily classifiable.

- **Constrained** – The constrained data condition limits the incident language material used to train/adapt the tested technology to only those distributed according to Section 5 (IL Data) and

[1] http://www.nist.gov/itl/iad/mig/lorehlt16.cfm

[2] This task is for year 1 only. In subsequent years (2+), the task will be Entity Discovery and Linking (EDL)

Section 6 (Native Language Informants). No other incident language materials, i.e., parallel text, speech corpora, etc. are permitted but knowledge gained from the Native Language Informant is permitted. Prior to the evaluation period, which begins with the announcement of the IL, teams can assemble multilingual resources/technologies/etc. to use during the evaluation so long as they are multilingual-focused in nature. Serendipitous included incident language data in a multilingual system is allowed. The use of mono- and bi-lingual resources is allowed so long as they do not include the incident language. The Constrained training condition is **required for each task participated**.

- **Unconstrained** – The unconstrained condition removes the limitations of the constrained condition. Teams can use additional, publicly available, incident language materials obtained before or after the IL announcement from an epoch before or after the incident. Teams can use pre-existing, mono-lingual technologies for the incident language. Teams can use additional Native Language Informant time beyond the limits in Section 6. The teams must document the additional data and technologies in their system description. The unconstrained training condition **optional but encouraged**.

4 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, volunteer participants must register to participate and sign the license agreement which can be found on the NIST LoReHLT website.

5 Evaluation Data

5.1 Component Definition & Release Plan

All three evaluation tasks have the same data components and release plan. The LDC releases the Incident Language (IL) data and English Scenario Model in an encrypted format, and NIST releases the appropriate decryption key(s) at the appropriate stages. Participants must complete all three checkpoints for their submissions to be considered complete. 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 7 days after IL Announcement
  - Decryption key for **set 1** released 7 days after IL Announcement and after submission to Evaluation Checkpoint 1 made

- Evaluation Checkpoint 2
    - Train with data from **set 0** begins at IL Announcement
    - Train with data from **set 1** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
    - Evaluation Checkpoint 2 submission due 14 days after IL Announcement
    - Decryption key for **set 2** released 14 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** begins after the Evaluation Checkpoint 1 submission deadline and the team makes a submission
    - Train with data from **set 2** and **set S** begins after the Evaluation Checkpoint 2 submission deadline and the team makes a submission
    - Evaluation Checkpoint 3 submission due 30 days after IL Announcement

5.2 Data Description

The composition of the five datasets (**set 0, set 1, set 2, set S, and set E**) are listed in the table below. The sizes given are approximate, and "Kw" refers to multiples of 1000 words.

Set 0 – pre-incident epoch
Monolingual Source Text: • 100Kw newswire • 75Kw discussion forum/blog • 50Kw Twitter/SMS Parallel Text: • 100Kw newswire • 100Kw discussion forum/blog • 100Kw Twitter/SMS *300Kw comparable may be substituted for 100Kw parallel if parallel text is not available 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

6 Native Informant Resources

During the evaluation period, participants are allowed the use of a native informant in their system development. The LORELEI performers will be provided the native informant by their sponsor[3] through the data provider Appen. The native informant will be available remotely via telephone or internet connection. Volunteer participants, if they wish to use a native informant, have to supply their own at their own cost. It is up to the volunteer participants to determine how they communicate with their informant. However, consultation with the informant, by LORELEI performers and volunteer 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 probing of the evaluation data is 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 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)

7 Evaluation Protocol

7.1 Evaluation Account

All participants will be required to sign up for an evaluation account on the NIST LoReHLT evaluation web site in order to register for evaluation tasks and complete LDC data license agreements. Participants must complete the LDC data license agreement to receive the baseline training set as well as the

[3] No additional resources will be provided by the sponsor.

evaluation data. All data will be made available directly by the LDC. Participants will upload submissions through the evaluation web site and be able to view submission status and results as these features become available.

A link to the evaluation web site, along with further instructions, will be provided when registration opens. See section 12 for the schedule.

7.2 Submission Requirements

Each site must submit the constrained training condition for all three evaluation checkpoints. Each site can submit up to five (5) runs for each training condition at each checkpoint and must designate one as the primary run for cross system comparisons. Participants are required to submit a description of each of their systems.

Submission formats and naming conventions are to be determined and will be made available prior to the dry run.

8 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[4].
- The participant agrees to process at least the constrained training track for each of the selected tasks.
- The participant agrees to complete all three checkpoints to be considered a complete submission for each selected task and training track combination.
- The participant agrees to participate in the dry run exercise to ensure evaluation readiness.
- 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.

9 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.

[4] 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.

9.1 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 must refrain from publishing results and/or releasing statements of performance claiming winning or be perceived as a ranking amongst other participants.
- 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.
- 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.

10 Dry Run

All participants are required to participate in a dry run evaluation to demonstration evaluation readiness. The purpose of the dry run is to iron out all of the bugs in the evaluation pipeline, not to gauge any system's capability. The dry run will follow the exact protocol of the official evaluation except that participants have to complete only the first evaluation checkpoint instead of all three.

11 System Description

Each team is required to submit a system description of the system(s) used for its submissions. The format of the system description will be posted on the evaluation website.

12 Schedule

Milestone	***Date***
Evaluation plan published	Dec 11, 2015
Registration opens	Feb 19

6-month PI meeting in San Antonio, TX (LORELEI performers only)	Feb 24 – 26, 2016
Registration deadline (Volunteer participants)	May 2
Required dry run begins	Jun 2
Required dry run ends	Jun 8
Teams receive encrypted IL data	Jun 29
IL Announcement Decryption keys for **set 0** and **set E** distributed	Noon Jul 6
Evaluation Checkpoint 1 submissions due Decryption key for **set 1** distributed after submission made	Noon Jul 13
Evaluation Checkpoint 2 submission due Decryption key for **set 2** distributed after submission made	Noon Jul 20
Evaluation Checkpoint 3 submission due	Noon Aug 3
System description due	Aug 17
Human Assessments	TBD
Human Assessment results released	TBD
2-day NIST evaluation workshop in Stevenson, WA (Volunteer participants; optional for LORELEI performers)	Aug 28 – 29
3-day DARPA PI meeting in Oregon (LORELEI performers only)	Aug 30 – Sep 1
Meeting to discuss Human Assessment Results	~Nov

13 Machine Translation (MT) Evaluation Specifications

13.1 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.

13.2 Performance Measurements

BLEU and METEOR will be the primary metrics in Phase 1. BLEU and METEOR scores will be calculated at each checkpoint. Scoring will be done against four 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.

NIST will investigate additional automatic metrics, as well as human assessment approaches, geared towards measurement of successful translation of content.

13.3 Input and Output Format

NIST has developed a DTD which defines the structure of the XML documents used to format MT source, reference, and translation files. The formatting requirements for this component MT evaluation are similar to those of the recent OpenMT evaluation. The DTD can be found here: ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-lorelei-p1.dtd. NIST requires that all submitted translation files are well-formed and valid against this DTD. LDC will provide the data in LTF format conforming to the LTF DTD referenced. A conversion script will be provided to convert the data in LTF format into the MT evaluation format specified by NIST. Either file format may be used for processing the evaluation data, but the MT system output must be in the MT evaluation format specified by NIST.

Below are samples of the MT source, reference translation, and system translation files. The text of the segments has been replaced by English placeholders.

Sample MT source file, LTF format:

```
<?xml version="1.0"?>
<!DOCTYPE LCTL_TEXT SYSTEM "ltf.v1.2.dtd">
<LCTL_TEXT lang="URD" source_file="NW_JAN_URD_023735_20060331.rsd.txt"
           source_type="web_news" author="LDC" encoding="UTF-8">
  <DOC id="NW_JAN_URD_023735_20060331" lang="URD">
    <TEXT>
      <SEG id="NW_JAN_URD_023735_20060331-1" start_char="3"
           end_char="17">
        <ORIGINAL_TEXT>source segment1</ORIGINAL_TEXT>
        <TOKEN id="NW_JAN_URD_023735_20060331-1-1"
               start_char="3" end_char="8">source</TOKEN>
        <TOKEN id="NW_JAN_URD_023735_20060331-1-2" start_char="10"
               end_char="17">segment1</TOKEN>
      </SEG>
    <SEG id="NW_JAN_URD_023735_20060331-2" start_char="19"
         end_char="33">
      <ORIGINAL_TEXT>source segment1</ORIGINAL_TEXT>
      <TOKEN id="NW_JAN_URD_023735_20060331-2-1" start_char="19"
             end_char="24">source</TOKEN>
      <TOKEN id="NW_JAN_URD_023735_20060331-2-2" start_char="26"
             end_char="33">segment2</TOKEN>
    </SEG>
…
    </TEXT>
  </DOC>
</LCTL_TEXT>
```

Sample MT source file, NIST MT evaluation format:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE mteval SYSTEM "mteval-lorelei-p1.dtd">
<mteval>
  <srcset setid="NW_JAN_URD_023735_20060331" srclang="URD">
```

```
    <doc docid="NW_JAN_URD_023735_20060331" genre="NW">
      <seg id="NW_JAN_URD_023735_20060331-1">source segment1</seg>
      <seg id="NW_JAN_URD_023735_20060331-2">source segment2</seg>
      …
    </doc>
  </srcset>
</mteval>
```

Sample MT reference translation file:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE mteval SYSTEM "mteval-lorelei-p1.dtd">
<mteval>
  <refset setid="NW_JAN_URD_023735_20060331" srclang="URD"
trglang="ENG" refid="reference01">
    <doc docid="NW_JAN_URD_023735_20060331" genre="NW">
      <seg id="NW_JAN_URD_023735_20060331-1">reference segment1</seg>
      <seg id="NW_JAN_URD_023735_20060331-2">reference segment2</seg>
      …
    </doc>
  </refset>
  <refset setid="NW_JAN_URD_023735_20060331" srclang="URD"
          trglang="ENG" refid="reference02">
    <doc docid="NW_JAN_URD_023735_20060331" genre="NW">
      <seg id="NW_JAN_URD_023735_20060331-1">reference segment1</seg>
      <seg id="NW_JAN_URD_023735_20060331-2">reference segment2</seg>
      …
    </doc>
  </refset>
  <refset setid="NW_JAN_URD_023735_20060331" srclang="URD"
          trglang="ENG" refid="reference03">
    <doc docid="NW_JAN_URD_023735_20060331" genre="NW">
      <seg id="NW_JAN_URD_023735_20060331-1">reference segment1</seg>
      <seg id="NW_JAN_URD_023735_20060331-2">reference segment2</seg>
      …
    </doc>
  </refset>
  <refset setid="NW_JAN_URD_023735_20060331" srclang="URD"
          trglang="ENG" refid="reference04">
    <doc docid="NW_JAN_URD_023735_20060331" genre="NW">
      <seg id="NW_JAN_URD_023735_20060331-1">reference segment1</seg>
      <seg id="NW_JAN_URD_023735_20060331-2">reference segment2</seg>
      …
    </doc>
  </refset>
</mteval>
```

Sample MT system translation file:

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE mteval SYSTEM "mteval-lorelei-p1.dtd">
```

```
<mteval>
  <tstset setid="NW_JAN_URD_023735_20060331" srclang="URD"
          trglang="ENG" sysid="NIST">
    <doc docid="NW_JAN_URD_023735_20060331" genre="NW">
      <seg id="NW_JAN_URD_023735_20060331-1">system segment1</seg>
      <seg id="NW_JAN_URD_023735_20060331-2">system segment2</seg>
      …
    </doc>
  </tstset>
</mteval>
```

14 Topic Labeling (TL) Evaluation Specifications

As the details of the Topic Labeling task are still being discussed, the content of this section is forthcoming.

15 Named Entity Recognition (NER) Evaluation Specifications

15.1 Task Definition

Given a document in the incident language, an NER system is required to automatically identify and classify entity mentions into pre-defined entity types. Note only named mentioned are targeted. The entity types in LORELEI/LORE tasks are listed as follows: (To be aligned with LDC NER annotations, we are planning to follow their definitions, so the following definitions may subject to change. A pointer to LDC's annotation guidelines will be given later.)

- Person (PER): Person entities are limited to humans identified by name, nickname or alias.
- Geo-political Entity (GPE): GPE entities are composite entities, meaning there are several criteria that must be present to make something a GPE. GPEs consist of (1) a physical location, (2) a government, and (3) a population. All three of these elements must be present for an entity to be tagged as a GPE, as in: United States, China, Pennsylvania, Philadelphia
- Organization (ORG): Organization entities are limited to corporations, institutions, government agencies and other groups of people defined by an established organizational structure.
- Location (LOC): Location entities include names of politically or geographically defined places (cities, provinces, countries, international regions, bodies of water, mountains, etc.). Locations also include man-made structures like airports, highways, streets, factories and monuments.

Other types of named entities like events, animals, inanimate objects and monetary units will not be annotated.

15.2 Performance Measurements

Scoring metrics from TAC KBP2014/2015 tasks will be extended to the NER tasks. System output will be computed against the gold annotation output for precision (P), recall (R) and their balanced harmonic mean (F1). The official metric will be based on exact mention boundary matches. Specifically, we report these three metrics (P, R and F1) for strong_mention_match (exact match), and strong_typed_mention from TAC2015 EDL measurements. The detailed description of TAC EDL scoring metrics is in section 2.1.2 in the overview paper: http://nlp.cs.rpi.edu/paper/edl2014overview.pdf.

In addition to the exact match metric, we award systems for partial matches according to the degree of character overlap between system and key names. The partial match scoring algorithm has two parameters: the recall overlap strategy and the precision overlap strategy.

- The per-name recall score of a name in the answer key is the fraction of its characters which overlap with the system name set according to the recall overlap strategy parameter. For the "MAX" strategy, this will be the characters overlapping with the single system name with maximum overlap. For the "SUM" strategy, this will be the number of its characters which overlap with any system mention.
- The recall score for a system is the mean of the per-name recall scores for all names in the answer key.
- The per-name precision score of a name in the answer key is the fraction of its characters overlapped by the reference set, where ”overlapping" is determined by the precision overlap strategy in the same manner as above for recall.
- The precision score for a system is the mean of the per-name precision scores for all names in the answer key.

We will report scores for all four parameter combinations.

15.3 Input and Output Format

The input and output formatting requirements for NER evaluation will be very similar to the most recent TAC2015 EDL evaluation. For more details, please refer to TAC2015 EDL task guidelines Section 2.3: http://nlp.cs.rpi.edu/kbp/2015/kbp2015edl_taskspec_v1.1.pdf

15.3.1 Input

The input for NER is a set of Incident Language documents provided by LDC. For offset calculation and formatting, we follow TAC2015 EDL’s guidelines described in Section 2.3: http://nlp.cs.rpi.edu/kbp/2015/kbp2015edl_taskspec_v1.1.pdf

15.3.2 Output

An NER 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 TAC2014/2015 EDL, some fields will just be place holders as noted below. Using the same format eliminates needs for making changes to the scorer code. Besides, full EDL is expected in year 2 and beyond.

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: NIL (in future this is a place holder for reference KB link entity 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. Each confidence value must be a positive real number between 0.0 (exclusive, representing the lowest confidence) and 1.0 (inclusive, representing the highest confidence), and must include a decimal point (no commas, please). Up to five answers to a given query may be included in each submission. The main score for the task will use only the highest confidence answer for each query, selecting the answer that appears earliest in the submission if more than one answer has the highest confidence value.