Option Period 1 Evaluation Plan for the IARPA MATERIAL Program

(**MA**chine **T**ranslation for **E**nglish **R**etrieval of **I**nformation in **A**ny **L**anguage)

Revision History Highlighted version number indicates released to performers.			
Version Number	**Date**	**By**	**Description**
0.0.0	06/03/2019	Audrey Tong	First internal version
0.0.1	06/19/2019	Audrey Tong	● Addressed T&E's comments ● Clarified that teams can only "manually examine closed queries and their relevance annotations after E2E eval" only if the language is released by IARPA. IARPA reserves the rights to keep certain language(s) sequestered for regression testing. ● Revised description for system output format and submission file format to accommodate that text and speech will be scored separately
0.0.2	08/15/2019	Audrey Tong	Added clarification that confidence factors must be consistent (e.g., the value for a "Yes" decision should be greater than the value for a "No" decision).
0.0.3	08/20/2019	Audrey Tong	Updated schedule
0.0.4	08/27/2019	Audrey Tong	Added ASR regression test
0.0.5	09/06/2019	Ilya Zavorin Audrey Tong	● Contingency matrix renamed into confusion ● Sec 3: added confusion matrix calculations ● Table 1: added language names and beta for 2C ● Sec 4: significantly rewritten ● Sec 7.2.2: component1 and 2 ● Sec 8.2: only 1B as the regression test language ● Table 7: updated ● Fixed typo and expanded ASR regression test submission procedure
0.0.6	09/11/2019	Ilya Zavorin	● Added 2C to 8.2 ● Added the 100-word limit and Basecamp link to JSON schema to 7.2.2 ● Added to 7.2.2 reference to an URL link for the aesthetics spec
0.0.7	09/17/2019	Ilya Zavorin	● Corrected equation numbers and references in Sections 3 and 4 ● Edited Section 4 including adding eq 5
0.0.8	09/18/19	Ilya Zavorin	Added sentence to end of Sec 4 about using K=1 for 2B/2S
0.0.9	09/30/19	Ilya Zavorin Audrey Tong	● Edited bottom half of Sec 2 ● Added a sentence to end of Sec 4 ● Updated 2B/2S eval week schedule ● Addressed NIST WERB reviewers' comments

CONTENTS

1 Introduction

This document describes the specifications for the evaluation of the second period of the MATERIAL (MAchine Translation for English Retrieval of Information in Any Language) Program. A continuation of the base period (BP) that began in October 2017 and ended in March 2019[1], Option Period 1 (OP1) still has the same objective which is to develop methods to locate content in speech and text "documents" in low-resource languages using English queries and to display summaries in English that convey why the system thinks the documents are relevant to the queries.

Like the BP, the queries will be in English, the material to be searched will be in different languages, and the summaries must be in English[2]. However, unlike the BP the queries will not be contextualized by domain. Furthermore, in OP1 domain and language identification tasks will not be evaluated. However, there will be regression tests to track performance of the various aspects of the system. Those tests are detailed in this evaluation plan. Data releases and evaluation cycle will also be simplified. The program metric AQWV (Actual Query Weighted Value) will remain the primary metric. However, system performance for each document mode (text, speech) will be computed separately, and a weighted average of the two modes will form the final score. While the primary metric is unchanged for OP1, other metrics will be explored to see if they will yield further insights into system performance.

2 Evaluation Tasks

For OP1, the task is given a set of foreign language documents and English queries, retrieve documents that are relevant to each query (Cross Lingual Information Retrieval or CLIR part) and generate a summary in English for each document the system deems relevant to a query (Summary or +S part). Both parts (CLIR and +S) generate outputs that are evaluated and which together provide insight into the performance of the overall end-to-end (E2E) system. Please note that MATERIAL summaries are query-biased, i.e. the purpose of a summary is to convey to an English speaker relevance of the corresponding original document to the query. It is not an English summary of the entire original document.

3 AQWV Metric for CLIR

Each performer system will calculate and report a numerical score in the range [0,1] for every query-document pair. As described in Section 1.B.2.1 of the MATERIAL Broad Agency Announcement (BAA)[3], performers will choose a value for a detection threshold θ that will optimize system's performance in terms of the program metric described below. Given a MATERIAL query, all documents scored at or above the threshold value will be marked by the performer system as relevant to the query and all documents scored below will be marked as not relevant[4]. This threshold value must be consistent across all queries for a given submission.

---

[1] https://www.nist.gov/sites/default/files/documents/2019/05/28/material_eval_plan_v6.0.4.pdf

[2] Developers are free to use any techniques they wish, but in developing this evaluation plan we have considered that methods from cross-language Information Retrieval (CLIR), machine translation (MT), and summarization could provide a possible base for the development of successful novel approaches.

[3] https://www.iarpa.gov/index.php/research-programs/material/material-baa

[4] The detection threshold is envisioned as being used as a dial by the end-user of a MATERIAL system, to be adjusted depending on the user's preference for higher precision versus higher recall.

For a given MATERIAL query Q, let the number of MATERIAL documents that are relevant to Q be $N_{Relevant}$ and let the number of non-relevant documents to be $N_{NonRelevant}$. Let the total number of documents in the corpus be $N_{Total} = N_{Relevant} + N_{NonRelevant}$. For a given value of the detection threshold θ, let:

- X_1 be the number of *true positives*, i.e. relevant documents that a Performer Team's System (PTS) marked as relevant
- $X_2 = N_{Miss}$ be the number of *misses/false negatives*, i.e. relevant documents that the PTS did not mark as relevant
- $X_3 = N_{FA}$ be the number of *false alarms/false positives*, i.e. non-relevant documents that the PTS marked as relevant.
- X_4 be the number of *true negatives*, i.e. non-relevant documents that the PTS did not mark as relevant.

Then, $N_{Relevant} = X_1 + X_2$ and $N_{NonRelevant} = X_3 + X_4$ and we define the Query Value QV for query Q and detection threshold theta θ as

$$QV(Q,\theta) = 1 - [\, P_{Miss}(Q,\theta) + \beta\, P_{FA}(Q,\theta)\,] \qquad \text{(equation 1)}$$

where

- $P_{Miss}(Q,\theta) = \frac{N_{Miss}}{N_{Relevant}}$ is the probability of a missed detection error (i.e., the PTS failed to find a relevant document),
- $P_{FA}(Q,\theta) = \frac{N_{FA}}{N_{NonRelevant}} = \frac{N_{FA}}{N_{Total} - N_{Relevant}}$ is the probability of a false alarm error (i.e., the PTS retrieved a non-relevant document as relevant),
- β is defined as a constant a-priori so that all systems will optimize their performance in the same P_{Miss} vs. P_{FA} tradeoff space. For OP1, β has the following value:

Language	CLIR	CLIR+S
2B (Lithuanian)	40	40
2S (Bulgarian)	40	40
2C (to be announced in January 2020)	40	40

Table 1: β for each language and task for OP1.

Also, the confusion matrix for the response of the PTS to a single Q is:

		Performer Team's System (CLIR/E2E)	
		R (Relevant)	N (Not Relevant)
Answer Key	R (Relevant)	X_1	X_2
	N (Not Relevant)	X_3	X_4

Table 2: Confusion matrix.

And equation 1 can be rewritten as

$$QV(Q,\theta) = 1 - \left(\frac{X_2}{X_1+X_2} + \beta\frac{X_3}{X_3+X_4}\right) \qquad \text{(equation 1')}$$

All queries will be weighted equally regardless of their respective $N_{Relevant}$[5]. We define the Query Weighted Value for the full set of queries as

$$QWV(\theta) = \frac{\sum_{i=1}^{NQ} QV(Q_i,\theta)}{NQ} \qquad \text{(equation 2)}$$

where

- Q_i is a specific query
- NQ is the total number of queries
- QV is defined in equation 1

$AQWV(\theta)$ is the Actual Query Weighted Value which is $QWV(\theta)$ calculated for the system running at its actual decision threshold. The reader will note the following:

- $AQWV(\theta) = 1.0$ for a perfect system
- $AQWV(\theta) = 0.0$ for a system that puts out nothing (all misses, no false alarms)
- $AQWV(\theta)$ can go negative if excessive false alarms are returned
    - $AQWV(\theta) = -\beta$ if none of the documents that are actually relevant (according to the answer key) are returned (so that $P_{Miss} = 1.0$), while all the documents that are actually non-relevant (according to the answer key) are returned (so that $P_{FA} = 1.0$)

Because $P_{Miss}(Q,\theta)$ is undefined when Q has no relevant documents, a modified version of AQWV[6] will be calculated using P_{Miss} on queries with relevant documents and P_{FA} on all queries with the formula:

$$QWV_M(\theta) = 1 - \left(\frac{\sum_{i=1}^{NQ_{Relevant}} P_{Miss}(Q_i,\theta)}{NQ_{Relevant}} + \beta\frac{\sum_{j=1}^{NQ} P_{FA}(Q_j,\theta)}{NQ}\right) \qquad \text{(equation 3)}$$

where $NQ_{Relevant}$ is the number of queries with relevant documents. QWV_M is what the scoring server will report.

$AQWV(\theta)$ will be calculated separately for each document mode (text and speech). The final CLIR performance score will be an equal weighted average of the AQWV of the two modes.

[5] One can similarly define Document Value and Actual Document Weighted Value metrics by considering individual documents rather than queries, but we do not plan to calculate it.

[6] This version is the primary metric and will be referred to as Modified AQWV.

4 AQWV METRIC FOR E2E

When a system identifies a document as relevant to a query, it must then generate a textual evidence in English to indicate why system believes the document's content is relevant to the query. In this section we explain the formulation of AQWV for E2E. For a given query *Q,* let $X^{CLIR}{}_1$, $X^{CLIR}{}_2$, $X^{CLIR}{}_3$, $X^{CLIR}{}_4$ be the elements of the PTS confusion matrix at the CLIR stage, as defined in Section 3. The PTS generates a summary if it deems the document is relevant (so if it is a true positive or a false alarm). We will use human judges to assess the quality of the summary. Let K_h be the number of human judges used to assess the relevance of a single document to a query using the corresponding summary, and let K be the final number of relevance judgments for the query-document pair. We have two possible ways of using the judgments:

- Convert all binary human judgments into a single binary judgment. That is, take the set of K responses and under some decision rule annotate the corresponding document as either relevant or not relevant. In this case $K = 1$.
- Use the individual responses directly. That is, annotate each document as having some number of relevant judgments and some number of not relevant judgments. In this case $K = K_h$.

There are four possible cases:

- A true positive document (one of $X^{CLIR}{}_1$) is judged by a human as relevant (i.e. it stays a true positive)
- A true positive document is judged by a human as not relevant (i.e. it is *reclassified* as a miss)
- A false alarm document (one of $X^{CLIR}{}_3$) is judged by a human as relevant (i.e. it stays a false alarm)
- A false alarm document is judged by a human as not relevant (i.e. it is *reclassified* as a true negative)

Note that human judgments are not collected for any of the $X^{CLIR}{}_2$ *or* $X^{CLIR}{}_4$ documents. For a given query Q , the full set of documents, and K final judgments per query-document pair, let:

- r_1 be the total number of judgments reclassifying true positives to misses, with $0 \le r_1 \le K\, X^{CLIR}{}_1$
- r_2 be the total number of judgments reclassifying false alarms to true negatives, with $0 \le r_2 \le K\, X^{CLIR}{}_3$

Then the elements of the PTS confusion matrix at the E2E stage can be calculated as follows:

- $X^{E2E}{}_1 = K\, X^{CLIR}{}_1 - r_1$
- $X^{E2E}{}_2 = K\, X^{CLIR}{}_2 + r_1$
- $X^{E2E}{}_3 = K\, X^{CLIR}{}_3 - r_2$
- $X^{E2E}{}_4 = K\, X^{CLIR}{}_4 + r_2$

QV_{E2E} can then be calculated from these using equation 1' as

$$QV_{E2E}(Q,\theta) = 1 - \left(\frac{X^{CLIR}{}_2 + r_1/K}{X^{CLIR}{}_1 + X^{CLIR}{}_2} + \beta\frac{X^{CLIR}{}_3 - r_2/K}{X^{CLIR}{}_3 + X^{CLIR}{}_4}\right) \qquad \text{(equation 5)}$$

2B/2S will be evaluated mostly like 1S so we will use the $K = 1$ approach but will consider the $K = K_h$ alternative for the surprise language evaluation. As with the CLIR score, we will calculate separate E2E

scores for speech and text modes using the Modified AQWV formulation, as well as an equally weighted average of the two.

5 Data Resources

NIST will release various data packs to performer teams during the program period for system development and testing. The data packs are described below while their distribution timeline is given in Section 9.

5.1 Build Packs

Performers will receive build packs for Automatic Speech Recognition (ASR) and Machine Translation (MT) training. There will be approximately 50 hours of audio for ASR (with 40/10 training/development recommended division) and 800k words of bitext for MT training. Performers may wish to use some of the build-pack transcribed audio and bitext for Dev purposes (e.g., doing deleted interpolation or n-fold cross-validation).

These build packs will consist of the following:

- Language-specific peculiarities and/or language specific design document(s) which contains information on the language:
  - What family of languages it belongs to
  - Dialectal variation
  - Orthographic information (including notes on any encodings that occur in our datasets)
    - Information on the character set
    - For a language written in a non-Latin character set, a transliteration into Latin characters
- Files of transcribed conversational audio in that practice language
  - The directory structure of the build pack will identify some of this as a Dev[7] set, but performers are free to re-partition this data in any way desired
- Conversational audio: some in 8-bit a-law .sph (Sphere)[8] files and some in .wav files with 24-bit samples
- The 800k words of bitext (sentences in the language and corresponding English translations)
  - We anticipate providing source URLs but probably little or no other metadata

5.2 Document Packs

The document packs contain six genres of "documents" listed in Table 3. Some metadata including the genre information will be provided in the document packs. Text files will be in UTF-8 .txt file format, and speech files will be in .wav file format.

The volume of text (number of documents as well as number of words) is expected to be substantially larger than the volume of speech. Perhaps ¾ of the documents will be text. Because perhaps ¼ of the

[7] Although somewhat similar in purpose, this Dev set (designed specifically to test and tune ASR models) is different from the one described in Section 5.2.1 (designed to test and tune E2E systems).

[8] Some tools to manipulate NIST Sphere format are available at https://www.nist.gov/itl/iad/mig/tools. Basic information about the Sphere format can be found at https://www.isip.piconepress.com/projects/speech/software/tutorials/production/fundamentals/v1.0/section_02/text/nist_sphere.text

documents will be speech, performer teams will need ASR[9]. Likewise, performers' systems will have to adapt to new genres, which is a key challenge for the program.

Conversational Speech data will originate as two-channel audio and will be provided to performers as two-channel audio with the two channels temporally aligned. When any of that data is transcribed, the two channels will be transcribed separately, and then those two transcripts will be combined/interleaved into a single transcript that reflects the temporal alignment. Conversational Speech transcripts provided to performers (for example, in the Analysis Pack) will all be of that combined/interleaved form.

Mode	Genre	Abbreviation
Text	News Text	NT
	Topical Text	TT
	Blog Text	BT
Audio	News Broadcast	NB
	Topical Broadcast	TB
	Conversational Speech	CS

Table 3: Genres of MATERIAL documents and their abbreviations.

Speech data may have background speakers or music. We do not intend to transcribe what is clearly background speech, and we do not expect to score such background speech for retrieval or summarization.

There are three types of document packs: *Dev, Analysis,* and *Evaluation*. All three are drawn from the same data pool to form mutually exclusive sets. In BP, Dev and Analysis were selected such that they had similar domain distribution. However, in OP1, Dev and Analysis were chosen such that they would have similar probability of query relevance.

5.2.1 DEV

We will provide to the performers a Dev Dataset drawn from the same data pool as the Evaluation Dataset that performers can use for internal testing purposes. The Dev Dataset will consist of about 650 documents for each language and will be released in its entirety. The Dev Dataset will also include query relevance annotation.

5.2.2 ANALYSIS

We will provide to the performers an Analysis Dataset also drawn from the same data pool as the Evaluation Dataset that performers can use for error analysis. The Analysis Dataset will be similar size as the Dev Dataset and will be released in its entirety. The Analysis pack will include query relevance annotation as well as English translations and transcriptions of the audio documents. The Analysis Dataset was selected such that it would have a similar probability of query relevance as that of the Dev Dataset.

5.2.3 EVALUATION

Unlike the BP, the Evaluation Dataset will be released in a single pack, and there will be no distraction documents in extraneous languages. The Evaluation Dataset is not guaranteed to have the same query relevance probability as that of the Dev or Analysis Dataset.

[9] Audio data in the build packs released at each period's kickoff and in the Analysis Dataset will come with transcriptions, but transcriptions will not be provided for the evaluation data. Performers' systems must ingest audio speech data automatically.

5.3 Query Packs

The program queries will be distributed to performers in two packs for each language under test in OP1. The first query pack will contain *open* queries where performers can conduct any automatic or manual exploration or data harvesting activities on the open queries as long as they are documented and disclosed. The second query release will contain *closed* queries where performers are only allowed to submit to NIST for scoring their results produced against the Analysis, Dev, or Evaluation document packs. These results must be generated by their fully automatic E2E systems with no human in the loop.

Results on the open queries will not be counted toward the final AQWV.

Table 4 shows the minimum number of queries, per language, expected to be released at the two stages.

	Number of Queries
Query1 Pack (open)	300
Query2 Pack (closed)	1000

Table 4: Query release counts per language.

5.4 Data Usage Restrictions

This section describes the rules for document and query use. A released language is one for which query relevance annotations for the Dev and Eval partitions have been released to the performer teams after the final E2E evaluation for that language[10].

	Build	**Dev**	**Analysis**	**Eval**
Manually examine documents **before** the language is released	Yes	No	Yes	No
Manually examine documents **after** the language is released	Yes	Yes	Yes	Yes
Manually examine Q1 and relevance annotations on <document set>	-	Yes	Yes	No
Manually examine Q2 and relevance annotations before E2E eval	-	No	No	No
Manually examine Q2 and relevance annotations after E2E eval	-	Yes	Yes	Yes[11]
Automatic processing of all queries (Q1, Q2)	-	Yes	Yes	Yes
Mine vocabulary from documents and queries for MT/ASR development	Yes	No	No	No
Train MT/ASR models on languages currently evaluated from <document set>	Yes	No	No	No
Automatically extract and process vocabulary from documents and queries for IR and Summarization	-	Yes	Yes	Yes
Parameter tuning	Yes	Yes	Yes	No
Index data for automated modeling and E2E component algorithms	Yes	Yes	Yes	Yes
Use IR models built from Dev or Analysis	-	Yes	Yes	No
Build and apply cross-lingual training models from languages not currently evaluated	Yes	Yes	Yes	Yes
Score locally (AQWV)	-	Yes	Yes	No[12]
Score locally (BLEU, WER)	Yes	No	Yes	No

[10] As of April 2019, only 1A (Swahili) and 1S (Somali) have been released.

[11] Only for the released languages. Please note that examining relevance annotations does not include examining the underlying documents. Relevance annotations of the eval set are released for CLIR research only. It is expected that Eval data will not be used for MT or ASR development.

[12] Unless the language has been released.

Table 5: Rules outlining what is allowable for query and document sets.

Performers should use the Dev Dataset to test their systems (one does not want to test on one's training data) and can also use the Dev Dataset as a held-out dataset to set the values of general system parameters.

Unlike the Dev Dataset, performers are free to examine the Analysis Dataset in detail, although it too should not be used as training data. We envision that the Analysis Dataset will help performers to do glass-box testing to understand why and how their systems generated particular outputs, including how their system made miss errors and false-alarm errors. Performers may use the Analysis 1 documents (i.e. the first pack of Analysis documents) and the open query relevance annotations (i.e. for the queries from first Query release pack) for "glass-box" analysis and parameter tuning of E2E systems, or their components, that are trained using other data. Performers should be mindful, however, of possible overfitting that may result from maximizing their components' performance on such a small set. Because transcriptions and translations for the Analysis Dataset will be provided, performers may calculate ASR WER (Word-Error-Rate) scores and MT BLEU[13] scores on the Analysis Dataset.

Evaluation Dataset is to be treated as a blind test.

Performer teams may mine the web for additional training and/or development test data. This paragraph is intended to clarify the restrictions mentioned at the top of page 11 of the BAA. Specifically, any such data harvested for training or development must be shared with the other teams after the end of the evaluation cycle in which it is first used (for example, after the CLIR+S end-to-end evaluation). In contrast, if teams purchase data, it must be shared with the other teams immediately (see the first full paragraph on page 11 of the BAA). In either case, as stated in the first full paragraph on page 11 of the BAA, teams must not hire native speaker consultants for data acquisition, system development, or analysis. For example, it is forbidden to use native speaker consultants to find or post-process any such data.

Performer teams may not use third-party commercial software in any part of their pipeline (e.g., transcription, translation, retrieval, summarization, language ID, data harvesting). Teams may use web-based MT software for translating a few words or phrases from the Analysis documents as a potential way to understand errors in their systems.

Performer teams may use the open queries in any way they wish but must document their usage. Performer teams must treat the closed queries as part of the blind evaluation set (no examination, no probing, no human in the loop). All closed queries remain closed for the duration of the program unless T&E specifies otherwise.

While data crawling may continue during a program evaluation, models applied to Eval data cannot be modified using any data collected by the crawling during the evaluation period. All machine learning or statistical analysis algorithms should complete training, model selection, and tuning prior to running on the Eval data. With a single exception[14], this rule does not preclude online learning/adaptation during Eval data processing during an evaluation so long as the adaptation information is not reused for subsequent runs of the evaluation collection. Performers must document the

[13] BiLingual Evaluation Understudy. See the original paper, "BLEU: a method for automatic evaluation of machine translation" at http://aclweb.org/anthology/P/P02/P02-1040.pdf

[14] Performers are not allowed to use text Eval data for adaptation of their ASR models to the speech Eval data.

ways their online learning/adaptation approaches incorporates information extracted from the Eval corpus.

No data or annotations may be distributed outside of the MATERIAL Program by participants.

5.5 Structure of Datasets Released to Performers

The following is a directory tree for a given dataset. Transcriptions, translations, and domain/query relevance annotations will only be provided for the Analysis Datasets.

```
IARPA_MATERIAL-<EvalPeriod>-<LangID>/
        README.TXT
        file.tbl
        index.txt
        <DatasetName>/
                audio/
                        src/
                                <DocID>.wav
                        transcription/
                                <DocID>.transcription.txt
                        translation/
                                <DocID>.translation.eng.txt
                text/
                        src/
                                <DocID>.txt
                        translation/
                                <DocID>.translation.eng.txt
```

<EvalPeriod> ::= { BASE | OP1 | OP2 | ... }

<LangID> ::= { 1A | 1B | 1S | 2B | 2S | ... }

<DatasetName> ::= { DEV | ANALYSIS | EVAL | ... }

<DocID> ::= MATERIAL_<EvalPeriod>-<LangID>_<DocumentNumber>

<DocumentNumber> is an uninformative 8-digit random number that we assigned to the document.

An example of a legal DocID would be `MATERIAL_BASE-1A_12345678`.

6 File Formats and Their Interpretation

NIST has implemented a scoring tool[15] to calculate scores for tasks listed in section 2. The scoring tool requires the system output and reference to follow certain formats. This section describes these formats.

File formats will be UTF-8 ASCII text, with fields on the same line separated by a tab character. Lines are to be terminated by a line feed character (no carriage-return), as is typical for Unix-based systems. Syntactically, a field may be empty.

[15] NIST will make public the scoring tool for performers to use at https://www.nist.gov/iarpa-material-machine-translation-english-retrieval-information-any-language-program.

6.1 Query Format

A query will consist of a query string (a word string).

Query ::= QueryString[,QueryString]

QueryString ::= [“, a-zA-Z0-9()+:<>[]_] (i.e., includes parentheses and square brackets)

There are four query types:

- **lexical** - requests the system to find documents that contain a translation equivalent of the query string. A translation equivalent should sound natural to a native speaker.
- **morphological** - requests the system to find documents that contain a translation equivalent matching the “marked” morphological properties of the query string.
- **full conceptual** - requests the system to find documents that contain the topic or concept of interest suggested by the query string.
- **EXAMPLE_OF** - requests the system to find documents mentioning an example of the query string.

A special query type called **conjunctive** is a logical *and* of any two query types listed above with the exception that both cannot be full conceptual and EXAMPLE_OF. Here are two examples, one lexical and one conjunctive query, respectively:

```
music
```

```
ebola,death
```

Refer to the MATERIAL Program Query Language Specification Document for a complete description of the query syntax including what is allowed and not allowed.

6.2 System Output Format

In OP1, text and speech will be scored separately. Therefore, systems are to output one file for text documents and one file for speech documents for each query. The name of these files must match the name of the corresponding reference files. The NIST scoring server will name the reference files using the query ID:

<QueryID>.tsv

For example:

```
query00043.tsv
```

The file content will have one line for every document from the corresponding speech/text document set along with the hard decision, confidence factor that the system assigned to that document for the given query, and optionally a metadata file to indicate information about the summary that the system generated. Those lines will be formatted as follows:

<DocID><tb><HardDecision><tb><ConfidenceFactor[16]>[<tb><Metadata File>]

Where:

<Metadata File> ::= <TeamID>.<SysLabel>.<QueryID>.<DocID>.json

An example for CLIR component only for the `query000043.tsv` would have 4 columns for each row:

[16] Confidence factors are specified in more detail in a later section of this evaluation plan.

```
MATERIAL_BASE-1A_12345678  Y  0.85
MATERIAL_BASE-1A_23456789  Y  0.840
MATERIAL_BASE-1A_34567890  Y  0.840
MATERIAL_BASE-1A_45678901  N  0.5
```

An example for CLIR and +S components for the `query000043.tsv` would have 5 columns for each row:

```
MATERIAL_BASE-1A_12345678    Y    0.85     FLAIR.MySystem1.query000043.MATERIAL_BASE-1A_12345678.json
MATERIAL_BASE-1A_23456789    Y    0.840    FLAIR.MySystem1.query000043.MATERIAL_BASE-1A_23456789.json
MATERIAL_BASE-1A_34567890    Y    0.840    FLAIR.MySystem1.query000043.MATERIAL_BASE-1A_34567890.json
MATERIAL_BASE-1A_45678901    N    0.5
```

The summary metadata file is currently going through some revisions. Please refer to the IARPA MATERIAL Program OP1 Summary Markup Specification document for what is being proposed. As soon as that information is finalized, it will be added to this section of the evaluation plan.

6.3 Reference Format

The reference files for the CLIR component on the scoring server will be named as:

<QueryID>.tsv

For example:

`query00043.tsv`

The format of the CLIR reference is similar to that of the CLIR system output format except no confidence factor field.

Assuming the dataset has 4 documents, a legal example of the CLIR reference file for `query000043` would be:

```
MATERIAL_BASE-1A_12345678  Y
MATERIAL_BASE-1A_52763409  Y
MATERIAL_BASE-1A_32198765  Y
MATERIAL_BASE-1A_98765432  N
```

6.4 Confidence Factors

For each query-document pair, the MATERIAL CLIR systems is required to give a confidence factor in the range 0.0 through 1.0, where 0.0 means “definitely non-relevant” and 1.0 means “definitely relevant.”

The confidence factor is to always have exactly one digit to the left of the decimal point, with at least one digit to the right of the decimal point, and no more than five digits to the right of the decimal point. The number of digits to the right of the decimal point need not be constant.

The confidence factor is *not* to be in any other floating point formats such as 5.0e-2. Examples of allowed confidence factors are:

```
0.0
0.5
0.54
0.54321
1.0
```

Examples of illegal confidence factors are:

`1` (must have a decimal point and at least one digit to the right of the decimal point)

`0.543211` (must have no more than five digits to the right of the decimal point)

Confidence factors of exactly 0.0 or exactly 1.0 have the same meaning across all systems. But this comparability *across systems* does not hold in between those values. More formally, for all confidence factors *cf* such that $0.0 < cf < 1.0$ there is *no* assumption that the confidence factors returned by one system are comparable to the confidence factors returned by another system. On the other hand, confidence factors returned by the *same system* on different queries for the same submission are assumed to be comparable; that is, the "Yes" decision threshold for one query is the same as that of another query. Confidence factors should be consistent which means a "No" decision should not have a higher value than a "Yes" decision.

7 Evaluation Scoring Server

NIST will provide an automated scoring server for the MATERIAL evaluation. Performer teams were given their own Google Drive (GD) where to deposit their submissions[17]. Performer teams must rename their submission to a particular naming convention so that the backend connecting to GD will know how to process their submissions.

Because in OP1 text and speech will be scored separately, performer teams must package the system output for text and for speech in separate submission files.

7.1 Submission Naming Convention

The naming convention for each submission is given below. The renaming script distributed by NIST can be used to generate this filename.

<SubmissionLabel> ::=
<TeamID>_<Task>—<SubmissionType>—<TrainingCondition>—<QuerysetID>—<SysLabel>_<EvalPeriod>—<LangID>—<NewDatasetName>_<Date>_<Timestamp>.tgz

where

<TeamID> = [a-zA-Z0-9]

<Task> ::= { CLIR | E2E | ASR }

<SubmissionType> ::= { primary | contrastive }

<TrainingCondition> ::= { unconstrained }, hard-coded[18]

<QuerysetID> ::= { QUERY1 | QUERY2 | NONE (for ASR) }

<SysLabel> ::= is an alphanumeric [a-zA-Z0-9] that performers assigned to the submission so they can keep track of which system output was submitted.

<EvalPeriod> = see section 5.5

<LangID> = see section 5.5

<NewDatasetName> := <DatasetName>-{SPEECH | TEXT} see section 5.5 for <DatasetName>

<Date> = <YYYYMMDD>

[17] The web version is no longer supported.

[18] At the end of a period when teams have shared all data resources, teams may be asked to run a "constrained" training condition utilizing the same shared resources to allow algorithmic comparison.

<Timestamp> = <HHMMSS>

For example:

```
NIST_CLIR-contrastive-unconstrained-QUERY2-mybestsystem_BASE-1S-EVAL-SPEECH_20
181113_225652.tgz
```

7.2 Packing System Output into Submission File

7.2.1 CLIR Submissions

System output files should be packed into a submission file. There should be no parent directory when the submission archive file is untarred. The renaming script previously distributed by NIST can be used to generate `<MySubmissionLabel>`. The tar command should be:

```
> tar zcvf <MySubmissionLabel>.tgz query*.tsv
```

The server will validate the submission file content to make sure the system output files conform to the format described in section 6.2.

7.2.2 E2E Submissions

A complete E2E submission will consist of a collection of individual directories each of which will contain all submission files corresponding to that query in a subfolder with the name <QueryID>, e.g.:

```
./query123/
      ./query123.tsv
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_23456789.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_34567890.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.component1.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_12345678.component2.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_23456789.component1.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_23456789.component2.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_34567890.component1.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_34567890.component2.jpg

./query45/
      ./query45.tsv
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_11223344.json
      ./FLAIR.MySystem1.query123.MATERIAL_BASE-1A_11223344.component1.png
```

For every conjunctive query, there will be 2 summary JPEG or PNG images per document (`component1` and `component2`). For a non-conjunctive query, there will be 1 summary JPEG or PNG image per document (`component1`). Up to 100 words per query component will be allowed, as specified in the `"eng_content_list"` element of the JSON schema[19]. Rendered summaries need to adhere to the aesthetic spec [20] that was designed to normalize basic elements of *form* of summaries rather than their *content*. A single zipped TAR `<MySubmissionLabel>.tgz` that will contain all query subdirectories.

[19] https://3.basecamp.com/3910605/buckets/5948786/uploads/1827264621

[20] https://3.basecamp.com/3910605/buckets/5948786/uploads/1803221431

The renaming script previously distributed by NIST can be used to generate `<MySubmissionLabel>`. The query-specific directories `<QueryID>)` will be collected together as follows:

```
> tar zcvf <MySubmissionLabel>.tgz *
```

8 Regression Tests

Occasionally during the program period, performer teams may be asked to reprocess previous evaluation data to track performance over time. Two regression tests are planned for OP1.

8.1 CLIR Regression Test

Performer teams will be asked to reprocess the 1B (Tagalog) evaluation data for the CLIR task. Teams will make only one submission for text and one for speech using the same system output format and submission protocol as the OP1 main evaluation. Please see the Schedule section for timeline.

8.2 ASR Regression Test

Performer teams will be asked to reprocess the non-distraction audio portion of the evaluation datasets of a subset of program languages and produce transcripts of the audio. WER will be calculated using NIST sclite scoring software[21]. There will be two regression tests: for 1B in December, 2019 and for 1B and 2C in July, 2020 (see Table 6).

8.2.1 ASR System Output Format

ASR system output will follow NIST ctm format. As described in the NIST sclite documentation, the ctm file format is a concatenation of time mark records for each word in each channel of a waveform. Each field in the record is separated by a space, and the records are separated with a newline. Each word must have a waveform id, channel identifier, start time, duration, and word token. Optionally a confidence score can be appended for each word. Each record follows this format:

CTM :== <F><sp><C><sp><BT><sp><DUR><sp>word[<sp><CONF>]

Where :

- <F> is the waveform base filename. NOTE: no pathnames or extensions are expected.
- <C> is the waveform channel. Either "A" or "B". The text of the waveform channel is not restricted by sclite. The text can be any text string without whitespace so long as the matching string is found in both the reference and hypothesis input files.
- <BT> is the begin time (seconds) of the word, measured from the start time of the file.
- <DUR> is the duration (seconds) of the word.
- <CONF> is an optional confidence score. Currently this field is not being used in sclite.

For example:

```
MATERIAL_BASE-1A_12345678 A 11.34 0.2 YES -6.763
MATERIAL_BASE-1A_12345678 A 12.00 0.34 YOU -12.384530
MATERIAL_BASE-1A_12345678 A 13.30 0.5 CAN 2.806418
MATERIAL_BASE-1A_12345678 A 17.50 0.2 AS 0.537922
```

[21] https://github.com/usnistgov/SCTK

```
:
MATERIAL_BASE-1A_12345678 B 1.34 0.2 I -6.763
MATERIAL_BASE-1A_12345678 B 2.00 0.34 CAN -12.384530
MATERIAL_BASE-1A_12345678 B 3.40 0.5 ADD 2.806418
MATERIAL_BASE-1A_12345678 B 7.00 0.2 AS 0.537922
:
```

8.2.2 ASR SUBMISSIONS

System output files should be packed into a submission file. There should be no parent directory when the submission archive file is untarred. The tar command should be:

```
> tar zcvf
NIST_ASR-contrastive-unconstrained-NONE-mybestsystem_BASE-1B-EVAL-SPEECH_20181
113_225652.tgz <DocID>.ctm
```

9 SCHEDULE (TENTATIVE)

During the evaluation week, teams can submit up to 5 submissions where one must be designated as *primary*. Primary submissions will be used to compare across teams and assessed by human judges in the case of E2E task. Submissions made during the evaluation week will not receive any score feedback.

In the case of CLIR and E2E, there should be one primary E2E following E2E file format and up to four contrastive CLIR following CLIR file format. There is no need to submit a CLIR primary since the CLIR primary results will be computed from the E2E primary.

Each submission will be validated prior to scoring. Only submissions that pass validation will count toward the submission limit. Submissions must follow the format given in the sections below.

Date	Event	Number of Submissions	Results Displayed
Apr 01, 2019	Virtual OP1 kickoff 2B/2S identity release 2B/2S build packs release		
Apr 01, 2019	2B/2S Q1/A/D source release 2B/2S Q1/A/D annotation release		
May 07-08, 2019	PI meeting		
May 28, 2019	2B/2S Post-hoc submission		
July, 2019	1B CLIR regression test	1	no
Oct 11, 2019	2B/2S Q2/E source release (by 10am Eastern time)		
Oct 14-23, 2019	2B/2S Eval week	5[22]	no
Oct 28, 2019	2B/2S CLIR results release		
Nov/Dec, 2019	2B/2S E2E results release		

[22] During the evaluation week, teams can submit up to 5 submissions for each mode (text or speech) where one from each mode must be designated as primary. Primary submissions will be used to compare across teams and assessed by human judges. Submissions made during the evaluation week will not receive any score feedback.

Dec, 2019	1B ASR regression test	1	no
Dec 01-10, 2019	Site visits		
Jan 06, 2020	2C identity release 2C build pack release		
Jan 06, 2020	2C Q1/A/D source release 2C Q1/A/D annotation release		
Feb, 2020	MATERIAL PMR		
May 15, 2020	2C Q2/E source release (by 10am Eastern time)		
May 18-22, 2020[23]	2C Eval week	5	no
May 27, 2020	2C CLIR results release		
TBD	2C E2E results release		
Jun 15-19, 2010	Site visits		
Jul, 2020	1B and 2C ASR regression test	1	no
Jul 13, 2020	OP1 final report and deliverables submitted End of OP1		
Aug, 2020	MATERIAL PMR		
Aug 10, 2020	Notification of OP2 award		

Table 6: OP1 schedule and evaluation submission quota.

[23] 2C Eval week dates are tentative and depend on the MATERIAL conference workshop dates.