Option Period 2 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**
1.0.0	10/14/2020	Audrey Tong Ilya Zavorin	List of major revisions relative to the OP1 Eval plan doc : ● Modified Sections 8.2 and 8.3 ● Added Sections 8.4 and 8.5 ● Updated Section 7.1 to match the changes above ● Added links to various spec documents to Section 7.2.2 ● Added OP2 schedule
1.0.1	11/17/2020	Audrey Tong Ilya Zavorin	Edits accompanying the 3C kick-off : ● Updated the 3C section (green fill) of table 6 ● Added Section 5.3.1 to include how 3C AN/DEV was revised after original document partition ● Modified Section 5.4 to indicate starting with 3C and after performers can view DEV documents (first row of table 5) and can use AN/DEV for training (last sentence in section). ● Added Section 8.6 to indicate performers to submit contrastive submissions for 3C Q2/EVAL using only text settings and updated Section 9 to reflect this
1.0.2	1/28/2021	Ilya Zavorin	Modified dates of the final PI meeting and final reports in Sec 9
1.0.3	3/26/2021	Ilya Zavorin	Edits following the discussion about E2E metrics : ● Updated CLIR AQWV targets for 3B and 3C in Table 2 ● Modified E2E AQWV beta for 3B and 3C in Table 2 ● Edited the beginning and end of Section 4 to give rationale for a higher beta for 3B and 3C and add F1
1.0.4	5/13/2021	Ilya Zavorin	Updated release date for 3B eval Edits following MT out of domain eval discussion : ● Initial version of Sec 8.7 with a high-level description of the eval ● Modified the regression section of the schedule (dates TBD) ● Modified dates of 3B ASR and MT submissions

CONTENTS

1 Introduction

This document describes the specifications for the evaluation of Option Period 2 (OP2), the third and final period of the MATERIAL (MAchine Translation for English Retrieval of Information in Any Language) Program. OP2 still has the same overall program 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. OP2 retains many of the main changes in Option Period 1 (OP1) including queries are not contextualized by domain, domain and language identification are not evaluated, and performance on text and speech are calculated separately. However, OP2 has several notable differences including:

1. The text documents in the Analysis, Development and Evaluation packs were harvested from crawled sources with loosely defined genres: "formal" (mostly news sources), "informal" (mostly blogs), and "topical" (from the CommonCrawl archive). These documents went through automatic cleaning and filtering steps but no human vetting with the exception of those assigned to the Analysis set.
2. Similarly, the bitext portions of the build packs will differ from those in the previous phases of the program in size, sources and/or translation conventions.
3. Another change is in the query types. Some query types will be dropped from the evaluation and part-of-speech will be added to the semantic constraints.
4. The test cycle will also be different to accommodate additional tests to probe certain aspects of the system that will inform the Program Manager of possible future program ideas. Those tests are detailed in this evaluation plan.

2 Evaluation Tasks

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 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 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)[1], performers will choose a value for a detection threshold θ that will optimize the 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 system as relevant to the query and all documents scored below the threshold will be marked as not relevant[2]. This threshold value must be consistent across all queries for a given submission.

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

[2] 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 system marked as relevant
- $X_2 = N_{Miss}$ be the number of *misses/false negatives*, i.e. relevant documents that the system did not mark as relevant
- $X_3 = N_{FA}$ be the number of *false alarms/false positives*, i.e. non-relevant documents that the system marked as relevant.
- X_4 be the number of *true negatives*, i.e. non-relevant documents that the system 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 system 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 system 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. The value for β is given in table 2.

Also, the confusion matrix for the response of the system to a single Q is given in table 1:

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

Table 1: Confusion matrix.

And equation 1 can be rewritten as

$$QV(Q,\theta) = 1 - \left(\frac{X_2}{X_1+X_2} + \beta\frac{X_3}{X_3+X_4}\right) \qquad \text{(equation 2)}$$

All queries will be weighted equally regardless of their respective $N_{Relevant}$[3]. We define the Query Weighted Value for the full set of queries as

$$QWV(\theta) = \frac{\sum_{i=1}^{NQ} QV\ (Q_i, \theta)}{NQ} \quad \text{(equation 3)}$$

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[4] will be calculated using P_{Miss} on queries with relevant documents and P_{FA} on all queries with the formula:

$$QWV_M(\theta) = 1 - \left(\frac{\sum_{i=1}^{NQ_{Relevant}} P_{Miss}(Q_i, \theta)}{NQ_{Relevant}} + \beta \frac{\sum_{j=1}^{NQ} P_{FA}(Q_j, \theta)}{NQ}\right) \quad \text{(equation 4)}$$

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).

OP2 has the following βvalue and target AQWV:

Language	CLIR		E2E	
	β	**Target AQWV (speech, text)**	β	**Target AQWV (speech, text)**
3S (Farsi)	40	0.6	40	0.6
3C (Kazakh)	40	0.6	600	TBD
3B (to be released on 4/16/21)	40	0.6	600	TBD

Table 2: β value and target AQWV for each language, task, and mode for OP2.

[3] 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.

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

4 METRICS 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 a system believes the document's content is relevant to the query. In OP2 we will be reporting two different E2E metrics: AQWV and F1.

Below we explain the formulation of AQWV for E2E, which remains the main E2E metric. For a given query Q, let X^{CLIR}_1, X^{CLIR}_2, X^{CLIR}_3, X^{CLIR}_4 be the elements of the system's confusion matrix at the CLIR stage, as defined in Section 3. The system 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[5]. 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 $0 \leq r_1 \leq K\, X^{CLIR}_1$
- r_2 be the total number of judgments reclassifying false alarms to true negatives, with $0 \leq r_2 \leq K\, X^{CLIR}_3$

Then the elements of the system's confusion matrix at the E2E stage can be calculated as follows:

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

[5] Details of this evaluation protocol will be documented separately.

- $X^{E2E}_4 = K\, X^{CLIR}_4 + r_2$

QV_{E2E} can then be calculated from these using equation 4 as

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

We are planning to run the OP2 evaluation with the same value of $K = 1$ that was used during OP1. As with the CLIR score, we will calculate separate E2E scores for speech and text modes using the Modified AQWV formulation.

Note that, as Table 2 shows, the value of βat the E2E stage was increased from 40 for the language 3S to 600 for the languages 3C and 3B. The reason for the change is to encourage performer teams to produce summaries that would yield higher values r_2(i.e. to reject more false alarms) while also keeping $r_1$low (i.e. to retain more true positives)[6].

We will also be computing the F1 metric at E2E as follows:

$$F1_{E2E}(Q,\theta) = \frac{2PR}{P+R} \qquad \text{(equation 6)}$$

where

- Precision $P = X^{E2E}_1/(X^{E2E}_1 + X^{E2E}_3)$
- Recall $R = X^{E2E}_1/(X^{E2E}_1 + X^{E2E}_2)$

5 Data Resources

NIST will release various data packs to performers 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. While the goal is to provide approximately 50 hours of audio for ASR (with 40/10 training/development recommended division) and 800k words of bitext for MT training similar to previous phases, the actual amount may be lower due to data availability. The bitext portions of the build packs may also differ from those in the previous phases in source distribution and/or translation conventions. Performers may wish to use some of the build-pack transcribed audio and bitext for development purposes (e.g., performing deleted interpolation or n-fold cross-validation).

The MT and ASR training resources 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

[6] See https://3.basecamp.com/3910605/buckets/5948786/messages/3512927973 for additional details

    - 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
- At most 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 speech and text documents like in previous phases. The speech documents remain similar with the same collection protocol and vetting. However, the text documents were harvested from crawled sources with loosely defined genres: "formal" (mostly news sources), "informal" (mostly blogs), and "topical" (from the CommonCrawl archive). While the text genres were mapped to the same abbreviations that were used in previous phases, they are not exactly the same due to how they were collected and vetted. The text documents had only automatic filtering and cleaning except for the analysis text documents. Table 3 lists the genre of speech and text documents. The speech documents are in .wav file format, and the text documents are in UTF-8 .txt file format.

The volume of text (number of documents as well as number of words) is substantially larger than the volume of speech. Because some documents will be speech, performers will need ASR[9]. Likewise, performers' systems will have to adapt to new genres, which is a key challenge for the program.

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.

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 the 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	Formal Text	NT
	Informal Text	BT
	Common Crawl	TT

[7] Although somewhat similar in purpose, this Development 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

[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 development or evaluation data. Systems must ingest audio speech data automatically.

Audio	News Broadcast	NB
	Topical Broadcast	TB
	Conversational Speech	CS

Table 3: Genres of MATERIAL documents and their abbreviations.

There are three types of document packs: *Analysis, Development,* and *Evaluation*. In BP, Development and Analysis were selected such that they had similar domain distribution, and in OP1, Development and Analysis were chosen such that they would have similar probability of query relevance; however, in OP2 the Analysis and Development document selection focuses on avoiding very short and very long documents for text and balancing acoustic conditions and sources for speech.

5.2.1 Analysis

Performers will receive an Analysis pack for error analysis. The Analysis set has a similar size as previous phases. The text documents came from the CommonCrawl archive. The Analysis pack includes query relevance annotation as well as English translations and transcriptions of the speech documents.

5.2.2 Development

Performers will receive a Development pack for internal testing purposes. The text document portion of the Development set is much larger and was crawled from online sources. The Development pack includes query relevance annotation.

5.2.3 Evaluation

Like OP1, there are no distraction documents in extraneous languages in the Evaluation pack.

5.3 Query Packs

The program queries[10] will be distributed to performers in two packs for each language under test in OP2. 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 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 approximate 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.

5.3.1 Repartitioning of 3C Query and Document Packs

To enable more robust development of MATERIAL systems for the speech modality, repartitioning of 3C query and documents was performed after the initial document partitioning and query development annotation efforts for this language were completed[11]. The goal of the repartitioning was to increase the number of relevant speech documents in Analysis and Development packs. The repartitioning started

[10] MATERIAL query typology is discussed in Zavorin, Ilya et al (2020).
[11] See https://3.basecamp.com/3910605/buckets/5948786/messages/2991503926 for additional details.

from the original set of Query1 queries (denoted here by *Q1*) and original sets of Analysis (denoted by *An*) and Development (denoted by *Dev*) documents for each modality. The speech documents from Dev were moved to the Analysis set resulting in a larger Analysis speech partition (denoted by *An+Dev*). Neither translations nor transcriptions of the Dev documents are provided. A subset of queries from Query2 that includes queries both with and without speech annotations were moved to Query1 (those are denoted by *Q2-speech*). Eval documents relevant to Q2-speech were also moved to the Development (denoted by *Dev'*) while relevant text documents remained in the Evaluation partition. Performers will use the following query-document combinations when reporting CLIR performance on Analysis and Dev:

- Text:
  - Q1 against An
  - Q1 against Dev
- Speech:
  - Q1 against An+Dev
  - Q2-speech against Dev'

5.4 Data Usage Restrictions

This section describes the rules for document and query use. An open language is one for which query relevance annotations for the Development and Evaluation partitions have been released to the performers after the final E2E evaluation for that language[12].

	Build	Dev	Analysis	Eval
Manually examine documents **before** the language is declared open	Yes	Yes[13]	Yes	No
Manually examine documents **after** the language is declared open	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[14]
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 Development 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[15]

[12] As of August 2020, 1A (Swahili), 1S (Somali), 2B (Lithuanian), 2C (Bulgarian), 2S (Pashto) are open languages.

[13] Starting with 3C and after.

[14] Only for the open 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.

[15] Unless the language has been declared open.

Score locally (BLEU, WER)	Yes	No	Yes	No

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

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

Unlike the Development 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[16] scores on the Analysis Dataset.

Evaluation Dataset is to be treated as a blind test.

Performers 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 performers after the end of the evaluation cycle in which it is first used (for example, after the E2E evaluation). In contrast, if performers purchase data, it must be shared with the other performers 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, performers 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.

Performers may not use third-party commercial software in any part of their pipeline (e.g., transcription, translation, retrieval, summarization, language ID, data harvesting). Performers 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.

Performers may use the open queries in any way they wish but must document their usage. Performers 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[17], 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 ways their online learning/adaptation approaches incorporate information extracted from the Eval corpus.

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

---

[16] 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

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

Starting with 3C, performers will be allowed to use the Analysis and Development sets for training, and will report the details of such use.

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 | 2C | 3B | 3C | 3S }

<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_OP2-3S_12345678`.

6 File Formats and Their Interpretation

NIST has implemented a scoring tool[18] 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 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.

6.1 Query Format

Query format remains the same in OP2 and consists of a query string (a word string).

[18] 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.

Query ::= QueryString[,QueryString]

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

Dropping morphological and EXAMPLE_OF, OP2 has two remaining basic 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. Example: `music`
- **conceptual**[19] - requests the system to find documents that contain the topic or concept of interest suggested by the query string. Example: `music+`

The special query type called **conjunctive** which is a logical *and* of any two basic query types and is limited to two lexicals in OP2. Example: `ebola,death`

Finally, part-of-speech is added to the query semantic constraint. It can occur alone or with another semantic constraint. Examples:

```
contest[n]
ring[n;evf:jewelry]
```

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

Like 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[20]>[<tb><Metadata File>]

Where:

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

An example for CLIR component only for the `query000043.tsv` would have 3 columns for each row:

```
MATERIAL_OP2-3S_12345678  Y   0.85
MATERIAL_OP2-3S_23456789  Y   0.840
MATERIAL_OP2-3S_34567890  Y   0.840
MATERIAL_OP2-3S_45678901  N   0.5
```

An example for CLIR and +S components for the `query000043.tsv` would have 4 columns for each row:

[19] In BASE and OP1, this was referred to as “full conceptual”.

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

```
MATERIAL_OP2-3S_12345678     Y     0.85      FLAIR.MySystem1.query000043.MATERIAL_OP2-3S_12345678.json
MATERIAL_OP2-3S_23456789     Y     0.840     FLAIR.MySystem1.query000043.MATERIAL_OP2-3S_23456789.json
MATERIAL_OP2-3S_34567890     Y     0.840     FLAIR.MySystem1.query000043.MATERIAL_OP2-3S_34567890.json
MATERIAL_OP2-3S_45678901     N     0.5
```

Contents of JSON files are described in Section 7.2.2.

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_OP2-3S_12345678  Y
MATERIAL_OP2-3S_52763409  Y
MATERIAL_OP2-3S_32198765  Y
MATERIAL_OP2-3S_98765432  N
```

6.4 Confidence Factors

For each query-document pair, the MATERIAL CLIR system 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. Performers were given their own team drive on Google Drive (GD) to deposit their submissions[21]. Performers must package their submission using the guidelines given in the sections below and deposit their submissions to their assigned team drive under the corresponding task directory so that the backend connecting to GD will know how to process their submissions. For example:

```
MATERIAL_Performer_FLAIR/CLIR/input
```

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> ::= { FLAIR | SARAL | SCRIPTS }

<Task> ::= { CLIR | E2E | ASR | MT | SLE }[22]

<SubmissionType> ::= { primary | contrastive }

<TrainingCondition> ::= { unconstrained }, hard-coded[23]

<QuerysetID> ::= { QUERY1 | QUERY2 | NONE }, use NONE if task is ASR or MT

<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> := { ANALYSIS | DEV | EVAL }--{ TEXT | SPEECH | SPEECH-REF-TRANSCRIPT}, use SPEECH-REF-TRANSCRIPT if MT is generated from the reference transcript

<Date> = <YYYYMMDD>

<Timestamp> = <HHMMSS>

For example:

```
NIST_CLIR-contrastive-unconstrained-QUERY2-mybestsystem_BASE-1S-EVAL-SPEECH_20181113_225652.tgz
```

[21] The web version is no longer supported.

[22] See Sections 7.2.1, 7.2.2, 8.2 , 8.3, and 8.4, respectively, for the submission requirements for these tasks.

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

7.2 Packing System Output into Submission File

7.2.1 CLIR Submission Guidelines

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 Submission Guidelines

A complete E2E submission will consist of a collection of individual directories. Each directory corresponds to a query, and inside each directory is a set of json files and their corresponding summary image component(s) for documents that the system deemed relevant. For example:

```
./query123/
      ./query123.tsv
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_12345678.json
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_12345678.component1.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_12345678.component2.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_23456789.json
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_23456789.component1.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_23456789.component2.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_34567890.json
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_34567890.component1.jpg
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_34567890.component2.jpg

./query45/
      ./query45.tsv
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_11223344.json
      ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_11223344.component1.png
```

For every conjunctive query, there will be 2 summary JPEG or PNG images per relevant 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[24]. Rendered summaries need to adhere to the aesthetic spec [25] that was designed to normalize basic elements of the *form* of summaries rather than their *content*. Contents of the `"eng_content_list"` element must adhere to the markup spec[26]. A single zipped TAR `<MySubmissionLabel>.tgz` that will contain all query subdirectories. 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 *
```

[24] https://3.basecamp.com/3910605/buckets/5948786/uploads/3076646738

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

[26] https://3.basecamp.com/3910605/buckets/5948786/uploads/3115724991

8 Additional Tests

During the program period, performers will be asked to perform additional tests to evaluate certain aspects of their systems. The subsections below describe these additional tests. Performers must package their submissions using the guidelines given below for each task and deposit their submissions to their assigned team drive under the corresponding task directory. For example:

`MATERIAL_Performer_FLAIR/ASR/input`

8.1 CLIR Regression Test

Performers will be asked to reprocess the Q2/2C (Pashto) evaluation data for the CLIR task. Performers only need to make one submission for text and one for speech using the same system output format and submission protocol as the main evaluation. Please see Section 9 for the timeline.

8.2 ASR Baseline/Regression Tests

Performers will be asked to run their ASR system on the 1B (Tagalog), 2C (Pashto), 3S (Farsi), 3B, 3C Analysis, Development, and Evaluation speech document sets. WER will be calculated using NIST sclite scoring software[27]. Performers are required to make only one submission per document set for each language under test. Please see Section 9 for the timeline.

8.2.1 ASR System Output Format

ASR system output will follow NIST CTM format. There should be one CTM file per document. 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. 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. For MATERIAL, we will use "1" for “inLine” and "2" for “outLine”.
- <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_OP2-3S_12345678 1 11.34 0.2 YES -6.763
MATERIAL_OP2-3S_12345678 1 12.00 0.34 YOU -12.384530
MATERIAL_OP2-3S_12345678 1 13.30 0.5 CAN 2.806418
```

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

```
MATERIAL_OP2-3S_12345678 1 17.50 0.2 AS 0.537922
:
MATERIAL_OP2-3S_12345678 2 1.34 0.2 I -6.763
MATERIAL_OP2-3S_12345678 2 2.00 0.34 CAN -12.384530
MATERIAL_OP2-3S_12345678 2 3.40 0.5 ADD 2.806418
MATERIAL_OP2-3S_12345678 2 7.00 0.2 AS 0.537922
:
```

8.2.2 ASR Reference Format

ASR reference will follow NIST STM format. There should be one STM file per document. As described in the NIST sclite documentation, the stm file consists of several fields to form a record. Each record is separated by a newline and contains: the waveform's filename, the channel identifier, the speaker's id, the begin time, the end time, and the transcript of the segment. Each record follows this format:

STM := <F> <C> <S> <BT> <ET> transcript . . .

where:

<F> The waveform filename. NOTE: no pathnames or extensions are expected.
<C> The waveform channel identifier. For MATERIAL, we will use "1" for "inLine" and "2" for "outLine".
<S> The speaker id, no restrictions apply to this name.
<BT> The begin time (seconds) of the segment.
<ET> The end time (seconds) of the segment.
transcript The transcript can take on three forms:

- a whitespace separated list of words
- empty string
- the string "IGNORE_TIME_SEGMENT_IN_SCORING". When the string "IGNORE_TIME_SEGMENT_IN_SCORING" is used as the transcript, the process which chops the hypothesis file to matching reference segments ignores all hypothesis words whose time-midpoints occur within the reference segment's beginning and ending time. The effect is to make these segment regions "out-of-bounds" for scoring, thus generating no errors from that time region.

For example:

```
MATERIAL_OP2-3S_12345678 1 MATERIAL_OP2-3S_12345678_1 11.34 17.50 HOW ARE YOU
MATERIAL_OP2-3S_12345678 2 MATERIAL_OP2-3S_12345678_2 1.34 7.00 I AM GOOD
:
```

8.2.3 ASR Submission Guidelines

System output files should be packed into a submission file. There should be no parent directory when the submission archive file is untarred. We expect at least one submission per document set for each language under test.

For example for 3S it would be:

```
NIST_ASR-primary-unconstrained-NONE-bestsys_OP2-3S-ANALYSIS-SPEECH_20200928_123456.tgz
NIST_ASR-primary-unconstrained-NONE-bestsys_OP2-3S-DEV-SPEECH_20200928_123456.tgz
NIST_ASR-primary-unconstrained-NONE-bestsys_OP2-3S-EVAL-SPEECH_20200928_123456.tgz
```

When uncompressed one submission, there is no parent directory.

```
MATERIAL_OP2-3S_12345678.ctm
MATERIAL_OP2-3S_87654321.ctm
:
```

8.3 MT Baseline/Regression Tests

Performers will be asked to run their MT systems on the 2C (Pashto), 3S (Farsi), 3B and 3C Analysis, Development, and Evaluation document sets. BLEU will be calculated only for the Analysis set since it is the only set with reference translation. NIST will use NIST's implementation of BLEU[28]. Performers are required to make only one submission per document set for each language under test. Please see Section 9 for the timeline.

8.3.1 MT System Output Format

MT system output is plain UTF-8 ASCII text with one line per segment. There should be one output file per document. NIST will convert to the XML format expected by the BLEU scoring script.

- For the Analysis text documents, performers should use the reference segmentation because NIST will score using the reference segmentation.
- For the Analysis speech documents, performers can use whatever segmentation that is natural to their systems. Performers will be asked to provide the segmentation information so that the MT can be matched to the source. NIST will merge all the segments and score at the document level.
- For the non-Analysis text and speech documents, performers can use whatever segmentation that is natural to their systems. Performers will be asked to provide the segmentation information so that the MT can be matched to the source.

8.3.2 MT Submission Guidelines

System output files and auxiliary information should be packed into a submission file. We expect at least one submission per document set for each language under test.

For example, for 3S it would be:

For MT from TEXT

```
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-ANALYSIS-TEXT_20200928_123456.tgz
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-DEV-TEXT_20200928_123456.tgz
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-EVAL-TEXT_20200928_123456.tgz
```

When uncompressed, say `DEV-TEXT` submission above, there is a parent directory with two subdirectories with files inside.

```
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-DEV-TEXT_20200928_123456/
```

`segmentation/` output from team's segmentation system with one segment per line. For ANALYSIS, we will assume the segmentation is the reference segmentation so the `segmentation/` directory is not needed.

```
MATERIAL_OP2-3S_12345678.seg.txt
MATERIAL_OP2-3S_87654321.seg.txt
:
```

`translation/` output from team's MT system using the segmentation in the directory above, one line per segment

```
MATERIAL_OP2-3S_12345678.mt.txt
MATERIAL_OP2-3S_87654321.mt.txt
```

[28] https://www.nist.gov/document/mteval-v14c-20190801.tar.gz

```
        :
```

For MT from SPEECH

```
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-ANALYSIS-SPEECH_20200928_123456.tgz
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-DEV-SPEECH_20200928_123456.tgz
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-EVAL-SPEECH_20200928_123456.tgz
```

When uncompressed, say `DEV-SPEECH` submission above, there is a parent directory with two subdirectories with files inside.

```
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-DEV-SPEECH_20200928_123456/
    transcription/  output from team's ASR system using team's own segmentation with one
                    segment per line.
        MATERIAL_OP2-3S_12345678.seg.txt
        MATERIAL_OP2-3S_87654321.seg.txt
        :
    translation/    output from team's MT system using the segments in the directory above, one
                    line per segment
        MATERIAL_OP2-3S_12345678.mt.txt
        MATERIAL_OP2-3S_87654321.mt.txt
        :
```

For MT from SPEECH-REF-TRANSCRIPT

```
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-ANALYSIS-SPEECH-REF-TRANSCRIPT_20200928_123456.tgz
```

When uncompressed the submission above, there is a parent directory with one subdirectory with files inside. Please note only ANALYSIS has its reference transcript released.

```
NIST_MT-primary-unconstrained-NONE-bestsys_OP2-3S-ANALYSIS-SPEECH-REF-TRANSCRIPT_20200928_1234
56/
    translation/    output from team's MT system using the reference segmentation, one line per
                    segment
        MATERIAL_OP2-3S_12345678.mt.txt
        MATERIAL_OP2-3S_87654321.mt.txt
        :
```

8.4 Source Language Evidence (SLE) Exercise

Performers will be asked to have their systems provide information from the source document to indicate evidence that the systems had used to determine the document as relevant for 3S (Farsi) Query 2 on Evaluation document set. Performers are required to make only one submission per document set for each language under test. Please see Section 9 for the timeline.

8.4.1 Source Language Evidence System Output Format

The system output should follow the source language evidence schema[29].

8.4.2 Source Language Evidence Submission Guidelines

A complete submission will consist of a collection of individual directories. Each directory corresponds to a query, and inside each directory is a set of files each containing source language evidence corresponding to a document that the system deemed relevant. For example:

```
./query123/
    ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_12345678.src.json
```

[29] The current version is v2.1 located at https://3.basecamp.com/3910605/buckets/5948786/uploads/1769764550

```
        ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_23456789.src.json
        ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_34567890.src.json

./query45/
        ./FLAIR.MySystem1.query123.MATERIAL_OP2-3S_11223344.src.json
```

A single zipped TAR `<MySubmissionLabel>.tgz` that will contain all query subdirectories. 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 *
```

8.5 CLIR Sprint Exercise

Upon release of language 3S, performers will be asked to do a "sprint" exercise to test their ability to develop a system to handle a new language under severely constrained "bare-bones" time and resource conditions. Performers only need to make one submission for text and one for speech using the same system output format and submission protocol as the main evaluation. The sprint exercise has the following restrictions, in addition to those listed in Section 5.4[30]:

- Performers are only allowed to use what has been provided by IARPA, namely, the 3S Build Pack (Section 5.1), the Analysis and Development document packs (Section 5.2) and the Query1 query pack (Section 5.3)
- The use of any external resources (e.g. dictionaries), datasets (such as those provided by LDC) or crawls of any sort is prohibited
- Also prohibited is the use of any pre-trained models, including:
    - Publicly available models like BERT
    - Those previously developed by members of a performer team under MATERIAL, BABEL, LORELEI or any other similar effort
- Performers are allowed to immediately start crawling English and foreign-language data to be used after the sprint is over, but this will be a completely decoupled effort:
    - Crawling will not be done by the same members of a performer team as those involved in the development of the sprint systems
    - No information from the crawls can be used in sprint system development

IARPA is not currently planning similar sprints for languages 3B or 3C.

8.6 Constrained Speech CLIR Submission for 3C

Performers will be asked to run their CLIR systems using only text settings on the 3C Q2/EVAL set as a contrastive submission[31]. Performers are required to make only one submission per document set. Please see Section 9 for the timeline.

8.7 MT Out-of-Domain Evaluations

The rationale behind this evaluation is to simulate a use case when a USG user has a "generic" MT model and they need to process new data that is out of domain, without spending a lot of effort collecting or annotating new in-domain data. We're focusing on 3S/Farsi since it's one of the later MATERIAL languages, so its technology is more advanced, and since it is a language of interest to the community.

[30] See the MATERIAL Basecamp thread https://3.basecamp.com/3910605/buckets/5948786/messages/2861900879 for additional details

[31] See https://3.basecamp.com/3910605/buckets/5948786/messages/3160371038 for additional details

Performers will be provided with a small amount of the target domain tuning/training data and asked to tune the models they developed for the November 2020 Farsi evaluation to a new data set. After a tuning period that would be constrained in both time and outside resources that would be used, performers will run their tuned systems on an evaluation set and submit their outputs to NIST for scoring.

9 Schedule (Tentative)

During the evaluation period, performers can submit up to 5 submissions where one must be designated as *primary*. Primary submissions will be used to compare across performers and assessed by human judges in the case of the E2E task. Submissions made during the evaluation week will not receive any score feedback. For sprint exercises and regression tests, only one submission is required.

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

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 above.

Date	Event	Number of Submissions	Results Displayed
July 31, 2020	Release of 3S: Lang ID, metric target value, BP, Q1/An/Dev, Analysis text translations		
July 31 - Aug 10, 2020	10-day constrained 3S sprint		
Aug 11-24, 2020	initial round of unconstrained development		
Aug 24-25, 2020	OP2 KO/PI meeting; Performers report on the 10-day sprint 3S results and any other latest results		
September 14, 2020	PMR		
November 4, 2020	release of 3S Eval data		
Nov 4-18, 2020	3S E2E eval		
November 18, 2020	Performers make 3S E2E submissions	5[32]	no
November 18, 2020	release of 3C		
November 25, 2020	Performers make 3S MT, ASR, source language evidence submissions		

[32] During the evaluation week, performers 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 performers and assessed by human judges. Submissions made during the evaluation week will not receive any score feedback.

December 2, 2020	T&E reports 3S CLIR AQWV, WER, BLEU scores		
January 4, 2021	T&E reports 3S E2E scores		
Feb 1-15, 2021	Tentative: (Virtual) Site Visits		
March 15, 2021	Tentative: PMR: report on 3S and first 3C results		
April 5, 2021	release of 3C Eval data		
Apr 5-16, 2021	3C E2E eval		
April 16, 2021	Performers make 3C E2E submissions	5	no
April 16, 2021	release of 3B		
April 21, 2020	T&E reports 3C CLIR AQWV scores		
April 23, 2021	Performers make additional 3C submissions: • ASR, MT • Source language evidence • Contrastive CLIR speech using text settings		
April 28, 2021	T&E reports additional 3C scores: • WER, BLEU etc • CLIR AQWV on contrastive speech using text settings		
May 16, 2021	T&E reports 3C E2E scores		
May 14, 2021	release of 3B Eval data		
May 14-28, 2021	3B E2E eval		
May 28, 2021	Performers make 3B E2E submissions	5	no
June 3, 2021	T&E reports 3B CLIR AQWV scores		
June 4, 2021	Performers make 3B ASR, MT submissions		
June 10, 2021	T&E reports 3B WER, BLEU scores		
TBD	release of 3S out-of-domain MT training/tuning data		
TBD	Performers tune their 3S MT systems to the out of domain data		

TBD	release of 3S out-of-domain MT eval data		
TBD	Performers make 3S out-of-domain submissions	1	no
August 2, 2021	T&E completes all evaluations		
Sep 13-14, 2021	Final PI meeting		
October 22, 2021	OP2 Final Report due from SCRIPTS, SARAL		
November 24, 2021	OP2 Final Report due from FLAIR		

Table 6: OP2 schedule and evaluation submission quota.

10 References

Zavorin, Ilya, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong and Richard Tong. 2020. Corpora for Cross-Language Information Retrieval in Six Less-Resourced Languages. C*ross-Language Search and Summarization of Text and Speech* (CLSSTS2020), Paris: European Languages Resources Association, pp. 7-13. https://bit.ly/302j7j0