

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE



MATERIAL: MAchine Translation for English Retrieval of Information in Any Language (Machine translation for English-based domain-appropriate triage of information in any language)

Dr. Carl Rubino
IARPA
September 27, 2016



Disclaimers

- This Proposers' Day Conference is provided solely for information and planning purposes.
- The Proposers' Day Conference does not constitute a formal solicitation for proposals or proposal abstracts.
- Nothing said at Proposers' Day changes the requirements set forth in a Broad Agency Announcement (BAA).



Proposers' Day Goals

- Familiarize participants with IARPA's interest in human language technology.
- Familiarize participants with IARPA's mission and how to do business with IARPA.
- Provide answers to participants' questions.
 - This is your chance to provide input to the program plan.
- Foster discussion of synergistic capabilities among potential program participants, i.e., facilitate teaming.
 - Take a chance – someone might have a missing piece of your puzzle.



Important Points

- Proposers' Day slides will be posted on iarpa.gov
- Please save questions for the end; write on notecards
- Posters are available for browsing during break/lunch
- Government will not be present during the poster/teaming session
- Discussions with PM allowed until BAA release
 - Once BAA is published, questions can only be submitted and answered in writing in accordance with the BAA guidance.



MATERIAL's Motivating Scenario





Navigating our multilingual world



ಅಂಕಾರಾ: ಟರ್ಕಿ ರಾಜಧಾನಿ ಅಂಕಾರದಲ್ಲಿ ಮಿಲಿಟರಿ ಪಡೆಯನ್ನು ಗುರಿಯಾಗಿರಿಸಿ ಬುಧವಾರ ಸಂಭವಿಸಿದ ಕಾರು ಬಾಂಬ್ ದಾಳಿಯಲ್ಲಿ 28 ಮಂದಿ ಸಾವಿಗೀಡಾಗಿದ್ದು, 61 ಮಂದಿಗೆ ಗಾಯಗಳಾಗಿವೆ. ಮಿಲಿಟರಿ ವಾಹನಗಳು ಹಾದು ಹೋಗುತ್ತಿರುವ ದಾರಿಯಲ್ಲಿ ಈ ಬಾಂಬ್ ಸ್ಫೋಟ ಸಂಭವಿಸಿದ್ದು, ಇದರ ಹಿಂದೆ ಯಾರ ಕೈವಾಡವಿದೆ ಎಂಬುದ ಸದ್ಯ ಪತ್ತೆಯಾಗಿಲ್ಲ ಎಂದು ಟರ್ಕಿ ಡೆಪ್ಯೂಟಿ ಪ್ರಧಾನಿ ನುಮಾನ್ ಕುರ್ತುಲ್ಮುಸ್ ಹೇಳಿದ್ದಾರೆ.

ಪಾರ್ಲಿಮೆಂಟ್ ಮತ್ತು ಮಿಲಿಟರಿ ಕೇಂದ್ರಗಳ ನಡುವೆ ಈ ಸ್ಫೋಟದ ಕಾರಣವೇನು ಎಂಬುದು ಅಧ್ಯಕ್ಷ ರೆಸೆಪ್ ತಯ್ಯಿಪ್ ಎಂದಿಲ್ಲದಿದ್ದರೂ ಅವರ ವಿರುದ್ಧ ಕ್ರಮ ಕೈಗೊಳ್ಳಲಾಗುವುದು ಎಂದು ಅಧ್ಯಕ್ಷ ರೆಸೆಪ್ ತಯ್ಯಿಪ್ ಎಂದಿಲ್ಲದಿದ್ದಾರೆ. ಈ ದಾಳಿಯ ಹಿಂದೆ ಜಿಹಾದಿ ಇಲ್ಲವೇ ಕುರ್ದಿಶ್ ಬಂಡುಕೋರರ ಕೈವಾಡವಿದೆ ಎಂದು ಟರ್ಕಿ ಸರ್ಕಾರ ಆರೋಪಿಸಿದೆ.

Military convoy

Car Bomb

Jihadist

New, unexpected languages do appear in an analyst's dataset which would require much time and investment to deploy automated methods for triage.



Working with a New Language

- Data in a new language may contain information critical for intelligence analysis but:
 - Many/most domain experts do not speak the language.
 - Few/no analysts speak the language.
 - No machine translation (MT) systems from that language to English are available to aid in analysis.
 - Only small amounts of new language to English bitext training data are available to build an MT-based system, and no domain-matched adaptation bitext is available to customize engines to support English-speaking analysts.
- If **Human Language Technology (HLT)** for the target language could be quickly deployed with output displayed in **English**, it would enable the domain experts to focus their efforts on the most relevant portions of the data



MATERIAL Goal

- Revolutionize multilingual triage by enabling rapid development of language-independent methods to field systems capable of fulfilling domain-specific cross-language information retrieval tasks over both text and speech data, with:
 - Limited bitext and transcribed speech training data
 - English domain-specific queries as input
 - English summaries of retrieved results as output
 - Methods for domain adaptation and portability to new languages
 - Assessment of the technology via a resonating end-to-end use case



The MATERIAL System

- An “English-in, English-out” information retrieval system that, given a domain-sensitive English query, will retrieve relevant data from a large multilingual repository and display the retrieved information in English as summaries that reflect the document relevance:











Query Format

- **Domain-specific (e.g., Agriculture, Government, Business, Entertainment)**
- **Address domain-restricted information need**

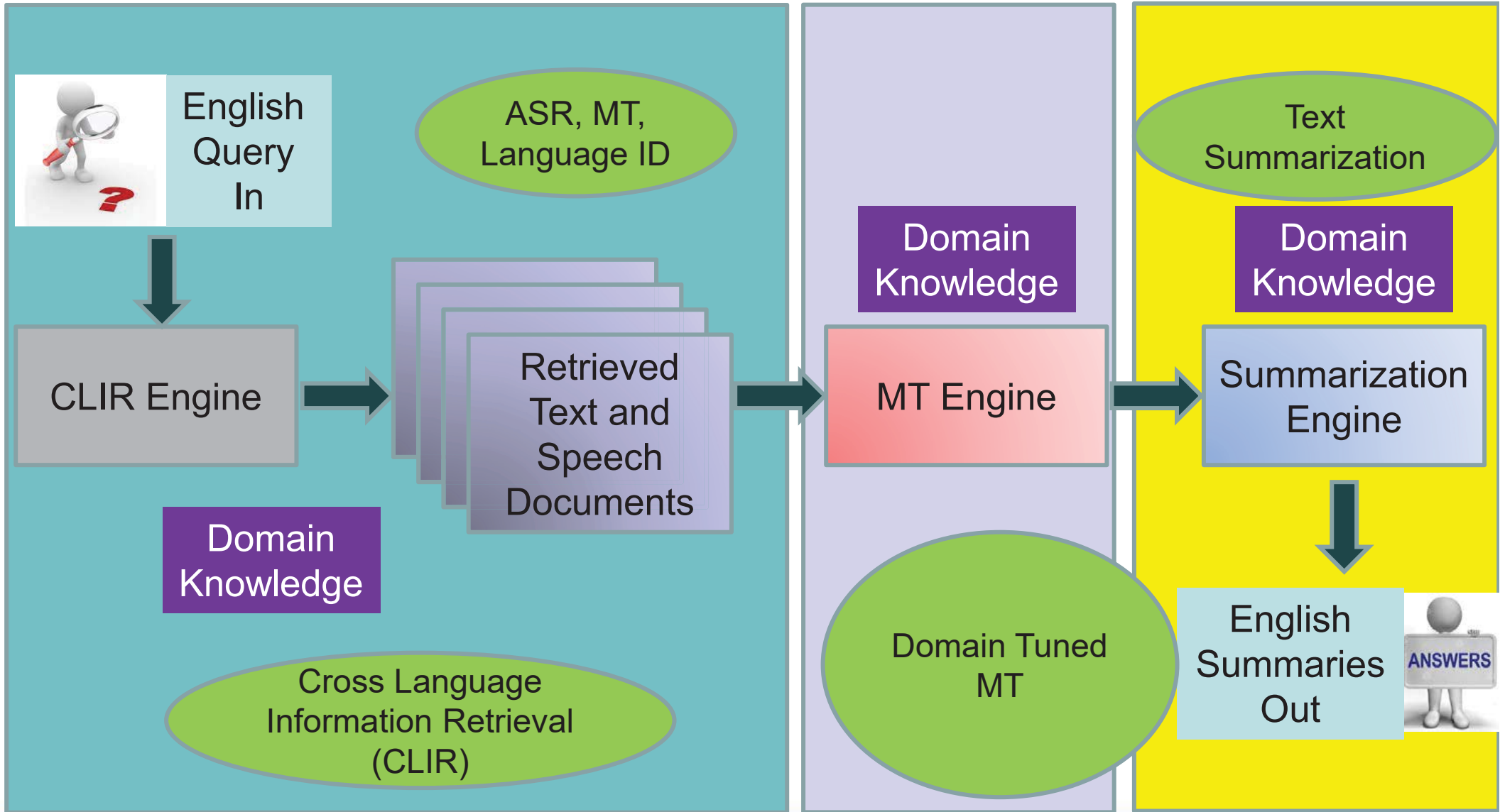
 “polio vaccine”
Domain: Government

Subject Domain

- | | | |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
|  |  | ...In response, the Armenian Ministry of Health urged all Syrian Armenians under age 15 to get the polio vaccination ... |
|  |  | ...Severe adverse reactions to this vaccine are rare.... |
|  |  | ...The oral vaccine was made by weakening the three strains of poliovirus that caused disease by growing them in monkey kidney cells... |



Technology Areas in Notional End-to-End System





Key Technical Challenges

- Techniques appropriate for a wide variety of languages
- Performance on formal and informal text and speech
- Development of new methods for domain adaptation without monolingual or parallel training data in that domain
- Limited development time
- Inter-language domain mismatches reflecting a cultural component

MATERIAL will emphasize minimizing training data needs for Automatic Speech Recognition (ASR) and MT.

MATERIAL will not provide domain adaptation data or domain annotated data. Performers will develop language-independent methods that are not data intensive.



MATERIAL Training Data

- Each language will be provided at kick-off to performers in a “pack” from the IARPA T&E Team that will contain training data for MT and ASR as well as relevant language information
- Speech Training data will include:
 - 35 - 45 hours of speech data with transcriptions in a normalized orthography with additional non-transcribed speech
 - Not all genres and domains will be present in the training data
- MT Training data will include:
 - 800K word bitexts in each program language, sentence-aligned
 - Not all genres and domains will be present in the training data
- Language information will include:
 - Description of the language (e.g., dialect regions, phoneme set definitions)
 - Basic information on dialects, spelling and encoding



Domains and Genres Used for Development and Evaluation

Program data will include formal and informal varieties of text and speech, including genres that are not present in the MT or ASR training data.

Mode	% Collect	Genre
Text	~ 75	News
		Topical
		Social Media
Speech	~ 25	Broadcast News
		Topical Broadcasts
		Conversation

Domains (Broad Subject Fields) may include: Agriculture, Science, Law and Order, Military, Sports, Politics, etc.



Test Structure Rationale

- T&E regimen designed to drive R&D towards the program goal, viz:
Language independent methods, tools, and technologies to provide **rapid-deployment** of **domain-adapted** MT for **low-resource** languages effectively integrated in a usable CLIR system
- So:
 - Multiple languages with varying characteristics
 - Only small amounts of IARPA-furnished bitexts for training
 - Domain contextualized queries
 - Decreasing lead-time for development & surprise language evaluation

	Base	Option 1	Option 2
Length (months)	18	16	12
# Practice Languages	2	2	3
Surprise Language Period (months)	6	4	1.5



MATERIAL Phase Structure

- Program Level Evaluations
 - CLIR evaluation (CLE)
 - Tests ability to find query-relevant source language “documents” in the appropriate domain of the query
 - CLIR + Summary evaluation (CLE+S)
 - Tests ability to find query-relevant source language “documents” in the appropriate domain
 - Tests ability to provide summaries that help a human reader identify relevant hits and filter out remaining irrelevant results
- Post-evaluation analysis
 - Bitexts and transcriptions will be available after each testing cycle to allow performers to correlate the performance of their ASR and MT systems with the program metrics. BLEU and WER scores will be provided.



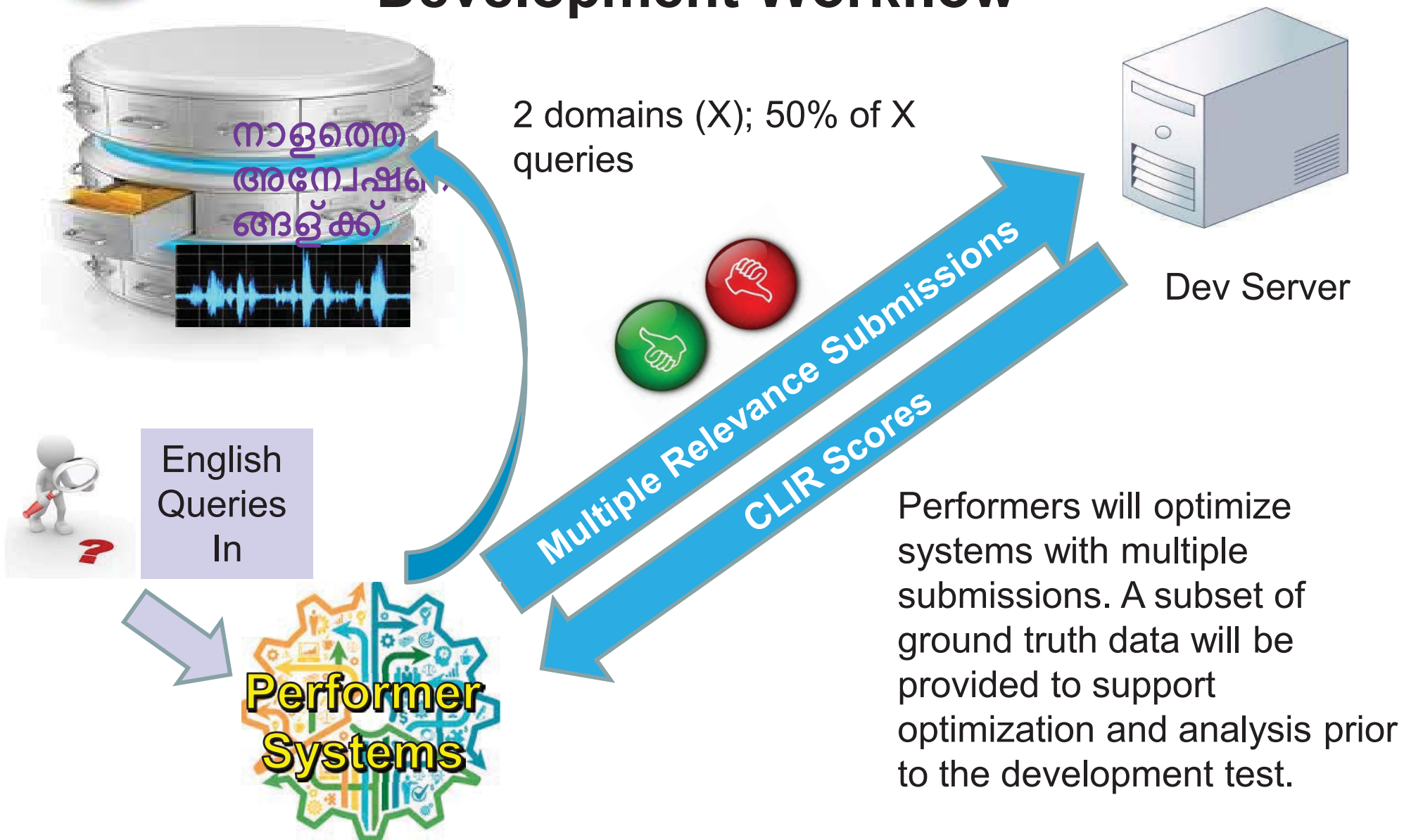
Query Release Schedule (per language)

Dataset Released	Epoch 1	Epoch 1+2	Epoch 1+2+3
# Domains	2	4	5
Query Set Introduced	X	X+Y	X+Y+Z
Practice Language Dev	50% Dev		
CLIR Evaluation		75% Dev 50% CLIR	
CLIR+Summary Evaluation			100% ALL

Performers will learn to handle new queries on new and old data, as well as old queries on new data.

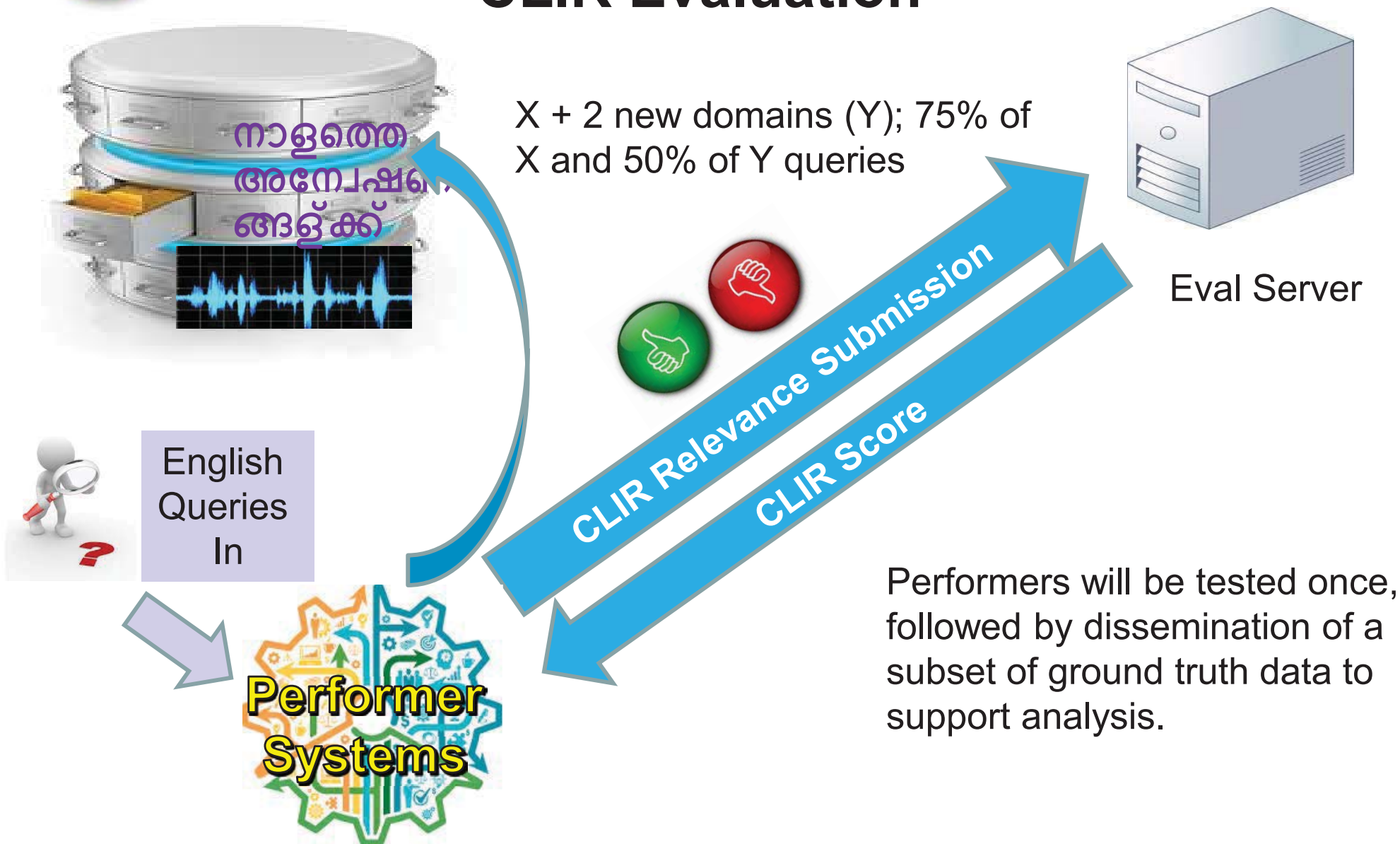


Development Workflow





CLIR Evaluation



നാളത്തെ അന്വേഷണങ്ങൾ



English Queries In

Performer Systems

X + 2 new domains (Y); 75% of X and 50% of Y queries

Eval Server

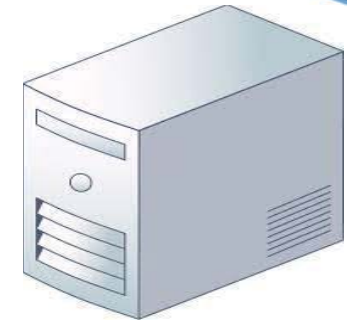
CLIR Relevance Submission

CLIR Score

Performers will be tested once, followed by dissemination of a subset of ground truth data to support analysis.



Final End-to-End Evaluation CLIR+S



Eval Server

Performers will be tested once, followed by dissemination of a subset of ground truth data to support analysis.

X+Y + 1 new domain (Z); All X, Y, and Z queries



Crowdsourced Judgments



നാളത്തെ അന്വേഷണങ്ങൾ



English Queries In

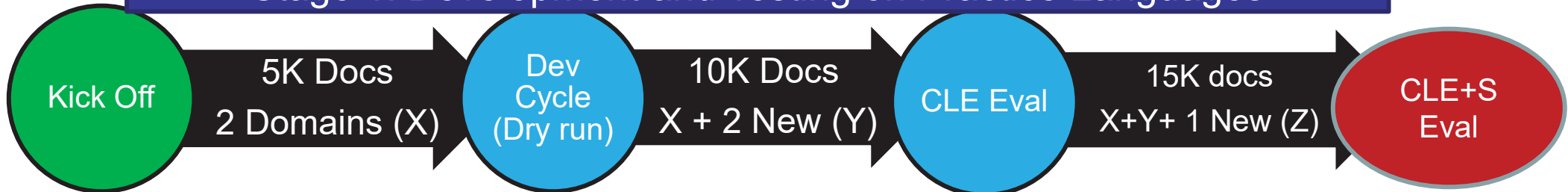




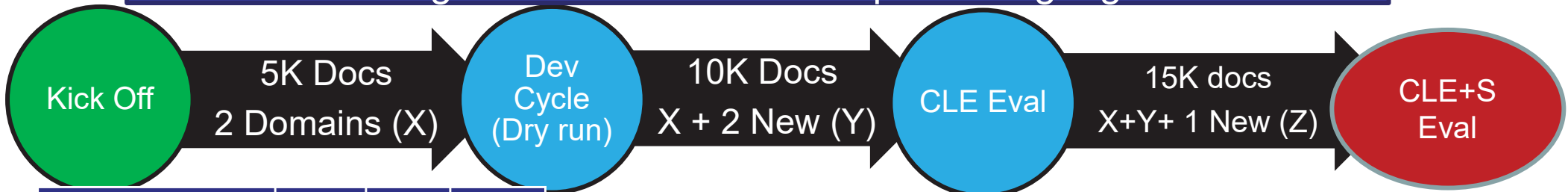
Program at a Glance

Training Data at Each Kickoff Period per language: 800K Words Bitexts; 35-45 Hours of Transcribed Audio

Stage 1: Development and Testing on Practice Languages



Stage 2: Evaluation on 1 Surprise Language



	Base	Opt 1	Opt 2
# Dev Languages	2	2	3
Phase Duration	18	16	13
Practice CLE	7	8	8
Practice CLE+S	10	10	9
Surprise CLE	13	12	10.5
Surprise CLE+S	16	14	11.5
Surprise Duration	6	4	1.5

Staging in of Queries:

Kickoff	Test 1	Test 2
50% X queries	75% X & 50% Y queries	100% X, Y & Z queries

10% responsive data translations provided at each stage for analysis that cannot be used for further training.

*Document sets and queries are reported per language



CLIR Detection Metric: AQWV Actual Query Weighted Value

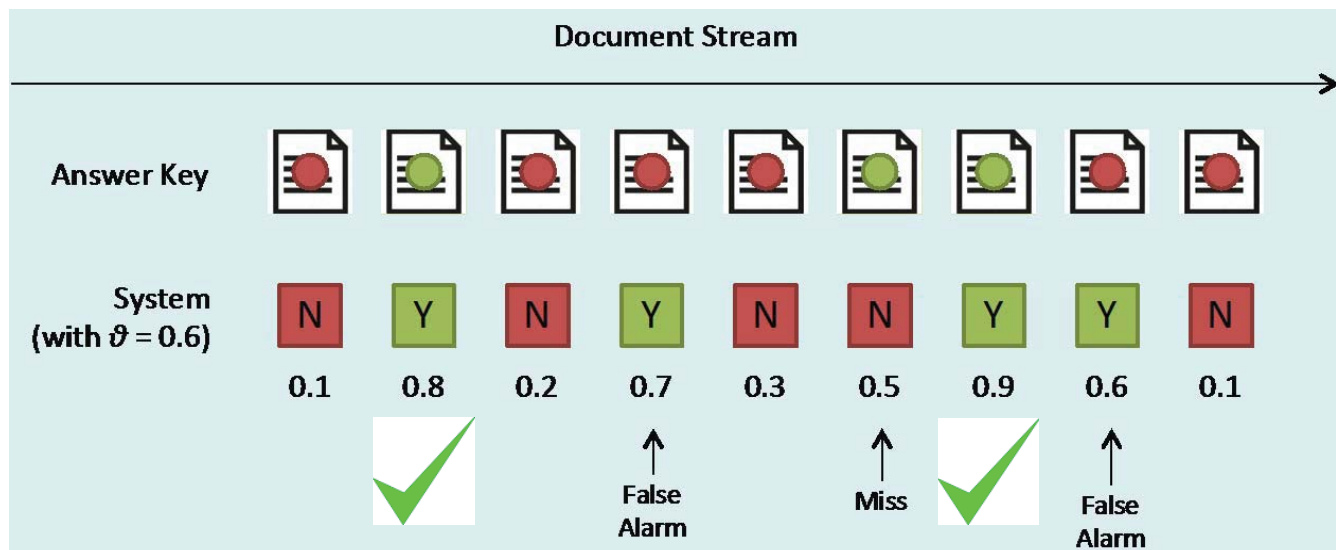
- All queries are treated equally (regardless of whether they generate single or multiple hits).
- Must be able to calibrate the metric against a baseline CLIR system that has two inputs: GOTS MT and human translation.
- Metrics will be reported to performers as an average over the set of queries (not individually for each query).
- Calculated as a representation of error rate taking into account probability of hits, false alarms, and the total number of responsive documents. These parameters will be set by T&E once the data are collected and evaluated.

Actual Query Weighted Value (AQWV)

Developers choose Θ , the detection threshold for their “Actual Decisions”, to optimize query-weighted value

- V is the *a priori* value (benefit) of a correct response
- C is the *a priori* cost of an incorrect response
- P_{rel} is the prior probability that a document is relevant to a query, e.g. 10^{-3}

$$Value_Q(\theta) = 1 - \underset{Q}{\text{average}} \left\{ p_{miss}(Q, \theta) + \frac{C}{V} (p_{rel}^{-1} - 1) \cdot p_{fa}(Q, \theta) \right\}$$



Developers will tune their systems to the threshold that maximizes the AQWV. Note that 1 is a perfect score; that is, error rate is zero.



Evaluating Summarization

CLIR Contingency Matrix

		Y	N
Key	Y	X_1 True Positive	X_2 False Negative
	N	X_3 False Positive	X_4 True Negative

$$QWV = 1 - \frac{X_2}{(X_1 + X_2)} - \beta \frac{X_3}{(X_3 + X_4)}$$

Crowd Summary Judgments

		Y	N
CLIR/ Key	Y/Y (= X_1)	A	B
	Y/N (= X_3)	C	D

CROWD-SOURCED JUDGMENTS:

Y/Y: Retrieved docs that are relevant

Y/N: Retrieved docs that are not relevant

A : # Relevant docs judged relevant

B : # Relevant docs judged non-relevant

C : # Non-relevant docs judged relevant

D : # Non-relevant docs judged non-relevant

A “perfect” summarization capability would hold B at zero and reduce C to zero



Evaluating Summarization (cont.)

CLIR Contingency Matrix

		Y	N
Key	Y	X_1 True Positive	X_2 False Negative
	N	X_3 False Positive	X_4 True Negative

$$QWV = 1 - \frac{X_2}{(X_1 + X_2)} - \beta \frac{X_3}{(X_3 + X_4)}$$

Crowd Summary Judgments

		Y	N
CLIR/ Key	Y/Y (= X_1)	A	B
	Y/N (= X_3)	C	D

A "perfect" summarization capability would hold B at zero and reduce C to zero

Summarization can reduce the false alarm rate ($C \leq X_3$) but cannot reduce the number of missed detections ($B \geq 0$)

End to End Contingency Matrix

		Y	N
Key	Y	A	$X_2 + B$
	N	C	$X_4 + D$

$$QWV = 1 - \frac{X_2 + B}{(A + X_2 + B)} - \beta \frac{C}{(C + X_4 + D)}$$



Summary

- Broad language portfolio:
 - Languages from a variety of language families (e.g., Afro-Asiatic, Niger-Congo, Sino-Tibetan, Austronesian, Altaic)
 - Mixed language typology (i.e., with different phonotactic, morphological, syntactic characteristics)
- Researchers will:
 - work with development languages to create new methods
 - be evaluated annually on a surprise language with development time and training data size constraints
- Evaluation:
 - On the set of development languages and the surprise language
 - Progress will be measured for:
 - [CLIR+S: AQWV](#) (Actual Query Weighted Value) metric and crowd sourced assessment of perceived relevance of delivered summaries
 - [MT, ASR](#): BLEU and WER scores for correlational analysis



Program Roles and Responsibilities

- **Performer R&D**
 - In Scope:
 - Novel methods for developing and adapting CLIR and MT of multi-lingual, multi-genre, multi-domain documents
 - Novel methods for developing summarization methods in English of retrieved documents for display
 - Novel use of machine learning, data resource gathering, and linguistics
 - Computational methods to reduce running time and memory footprint of models
 - Out of Scope:
 - Human User Interface
 - Image data
- **Government Support**
 - Government Furnished Information (GFI):
 - Acquire, organize and disseminate training data
 - Prepare and disseminate development, evaluation, and analysis data
 - Testing and Evaluation:
 - Evaluation framework to measure performer progress on practice and surprise languages
 - Development data to measure interim progress



Eligibility Information

- Other Government Agencies, Federally Funded Research and Development Centers (FFRDCs), University Affiliated Research Centers (UARCs), and any other similar type of organization that has a special relationship with the Government, that gives them access to privileged and/or proprietary information or access to Government equipment or real property, are not eligible to submit proposals under this BAA or participate as team members under proposals submitted by eligible entities.
- Non-US organizations and individuals may be able to participate.
 - Must comply with Non-Disclosure Agreements, Security Regulations, Export Control Laws, etc, as appropriate
 - Specific guidance for non-US participation will be provided in the BAA



Proposal Guidance

- Your proposal should include a full discussion of the technical approach that will be used to meet the program goals.
- Programmatic issues to be addressed in the proposal:
 - Your team's current technical capabilities
 - Key resources needed (not currently available to your team), to include capital equipment and special expertise (teaming will likely play an essential role in providing special expertise). The risk in acquiring these key resources, and mitigation strategies, should be indicated as well.
 - A teaming plan along with the roles and responsibilities of each member of the research team
 - End of phase and some intermediate milestones are set, but it is expected that other intermediate milestones that are on the critical path of the proposed approach will be offered.
 - A schedule of all milestones including a clearly charted description of the various risk mitigation strategies that will be undertaken to achieve program goals



Proposal Evaluation Criteria

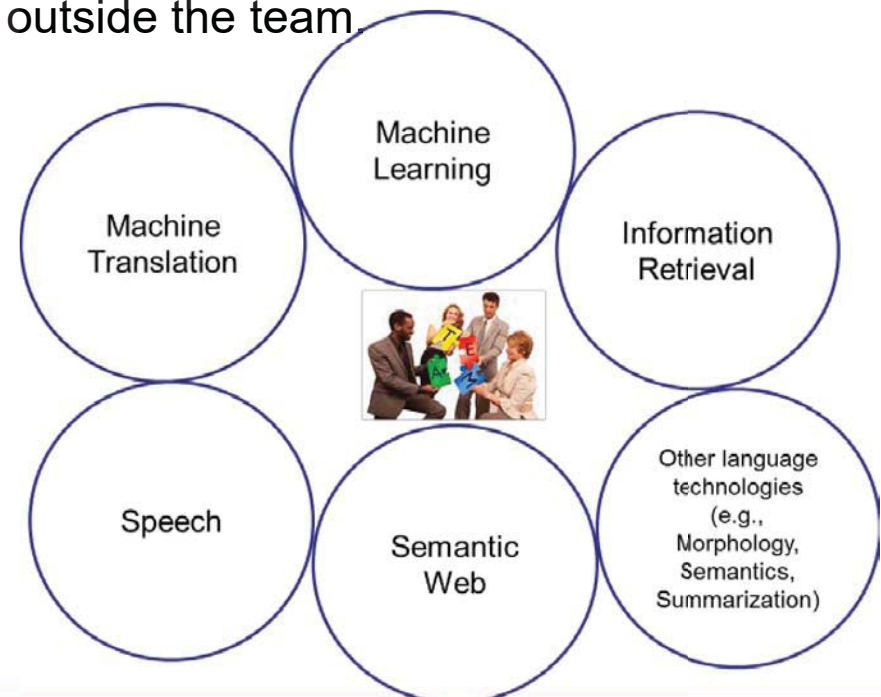
- Overall Scientific and Technical Merit
- Effectiveness of Proposed Work Plan
- Relevance to IARPA Mission and MATERIAL Program Goals
- Relevant Experience and Expertise
- Cost Realism

Evaluation criteria will appear in the BAA.

Teaming

- Because of the many challenges presented by this program, both depth and diversity will be strongly encouraged for overcoming these challenges.
 - Throughput: Consider all that you will need to do, all the ideas you will need to test. Make sure you have:
 - Enough people and expertise to do the job
 - Sufficient resources to follow critical path while still exploring alternatives – risk mitigation
 - Completeness: teams should not lack any capability necessary for success, e.g. should not rely on enabling technology to be developed outside the team.
 - Tightly knit teams
 - Clear, strong, management, single point of contact
 - No loose confederations
 - Each team member should be contributing significantly to the program goals.
Explain why each member is important, i.e. if you didn't have them, what wouldn't get done?

Can your team complete an evaluation well in the time frame required?





Additional Information

- Email dni-iarpa-baa-16-11@iarpa.gov with additional questions.
- MATERIAL BAA will be posted on FedBizOpps website (www.fedbizopps.gov).
- Q&As will appear after the BAA is posted. See http://www.iarpa.gov/solicitations_material.html.



Questions?