Decipher

Program Overview

Analysts often struggle to understand specialized language (e.g., slang, jargon, technical terminology) in various situations where dictionary definitions are not readily available, such as:

- Non-English research articles with novel technical concepts and terms
- Communications using slang
- Undefined acronyms used within limited contexts
- Web forum posts using coded language to mask illicit activities

In these and similar contexts, analysts dedicate significant effort to triaging content pertinent to their domain of interest and understanding the meaning of unknown or novel specialized terms used within that domain. Adding to this challenge, language can change quickly, and adversaries can rapidly adapt their language to evade detection.

The Decipher program's goal is to create capabilities to detect and define specialized language in diverse, multilingual text collections. In Decipher, "specialized language" refers to single or multiword expressions that are difficult to disambiguate (e.g. acronyms, terminology with multiple meanings), are used to deliberately obfuscate interpretation (e.g. coded language), or for which accurate translations or explanations are not readily available (e.g. emergent jargon or slang) to typical non-expert users. Decipher will flag and extract candidate specialized terms in a document or corpus, and then will provide candidate plain language translation and annotation of specialized terms, including jargon, slang, and acronyms. The program will also develop technology to annotate text with relevant social and contextual factors, capturing nuanced meaning, as well as to detect concept drift by monitoring use of terminology as it evolves over time.

Objectives

The Decipher capability will enable analysts to quickly interpret small text corpora of interest (e.g., a text message exchange containing limited messages). Additionally, it will enable the Intelligence Community (IC) to scour large, diverse collections, identify those which contain communications pertinent to IC missions, and furnish likely definitions in an easy-to-understand format.

Successful solutions must perform the following two automated tasks:

Task Area 1 (TA1): Detection: Find specialized language in a large corpus

For a given text collection, detect *specialized language*, which in the Decipher program refers to single or multiword expressions that are difficult to disambiguate (e.g. acronyms, polysemic terminology), are used to deliberately obfuscate interpretation (e.g. coded

DRAFT

language), or for which accurate translations or explanations are not readily available (e.g. emergent jargon or slang) to typical non-expert users.

A key challenge of this TA is defining a threshold between specialized and non-specialized language, as this may differ depending on the end user's knowledge of a domain as well as other factors. The Performer is encouraged to develop creative approaches and draw on techniques from the literature on lexical complexity prediction to address this challenge. Solutions should work on collections of multilingual content, especially high- and medium-resource languages.

Task Area 2 (TA2): Definition: Generate a glossary of specialized terms using limited context

Given a set of specialized language within a collection, generate a corresponding set of English-language definitions and plain language explanations or glosses for these terms with confidence scores.

Generated definitions should follow lexicographic best practices. Solutions should work on multilingual content and small collections of content containing limited attestations of a term. The TA2 capability should also accurately define common terms, though the focus of this TA is developing a tool that can generate definitions for specialized terms that are rare, unfamiliar, or previously unattested.

Additionally, Performers will develop a simple, user-friendly interface that presents the functionality of the technology for both TA1 and TA2, including specialized terminology definitions in glossary form and as annotations to a document.

Program Scope and Limitations

Proposed approaches shall explicitly address the following:

- Software Development: White papers shall describe the approach to software architecture, modularization, and integration. If the Offeror leverages backbone language models, the ability to "switch out" and retrain using future language model releases is highly desirable.
- Anticipated dataset development plans or preexisting datasets the Offeror plans to use in developing their system.
- Methods to address how the proposed solution will maximally leverage context to generate accurate definitions in situations where only a small number of attestations occur in a collection.

DRAFT

The following areas of research are out of scope for the Decipher program:

- Research that does not have strong theoretical and experimental foundations.
- Approaches that are unable to detect and generate definitions for novel future terms e.g., slang terms not attested until years after the completion of the program.
- Development of hardware.
- Approaches to develop new methods in cryptology (encoding) and cryptanalysis (decoding).
- Research that utilizes proprietary data that restricts the Government's ability to transfer the technology or data.
- Approaches that consist merely of integrating currently existing software.
- Approaches that only work on English-language materials.

Program Specific Terms/Definitions

Specialized Language: Single or multiword expressions that are difficult to disambiguate (e.g. acronyms, polysemic terminology), are used to deliberately obfuscate interpretation (e.g. coded language), or for which accurate translations or explanations are not readily available (e.g. emergent jargon or slang) to typical non-expert users)The Government, in consultation with the Test and Evaluation (T&E) Team may modify this definition to facilitate evaluation or optimize alignment with mission use cases.

Human or Animal Subject Research applicable to this Program

Performers are required to obtain human subjects research approval or a waiver from their organization's institutional review board where applicable.

Government Furnished Equipment/Information

No. While the Government will provide limited sample test data and information about the format of the test set, teams are expected to develop their own dev-test datasets.

Program Metrics

Performers will build and deliver prototype software systems that will be run and evaluated by the independent T&E Team in a closed compute environment on sequestered evaluation datasets. Testing protocols do not allow for expert operators, human-in-the-loop operation, or any operations not deemed "turnkey." The metrics below are preliminary examples and may be modified by the Government. Metric scores will be calculated by the T&E Team. The test set will contain novel specialized language developed to evaluate Performer system performance in a variety of conditions. This collection may contain, for example, a collection of text messages

DRAFT

containing a conversation about smuggling of weapons. Performers will not receive information about the contents of the collections or their topic domains.

The Government encourages proposers to recommend their own metrics to measure success. Good metrics for the purpose of this effort maximize interpretability, allow easy implementation and replicability, allow calculation of confidence bounds, furnish results that are comparable over time, and are actionable (i.e., inform changes in technical approach).

Detection

This metric focuses on the ability to accurately identify specialized language terms within a domain of user interest. Within the sequestered T&E environment, the Performer system will be provided with a prose description of the domain of interest at test time. The Government tentatively plans to use information extraction metrics (e.g., precision, accuracy, and F1) to evaluate how well Performer systems identify specialized language terms relevant within a given domain of interest. Performance in retrieving documents from a target domain of interest may also be measured relative to a baseline, which may be a general-purpose, state-of-the-art information retrieval capability.

Definition Generation

Metrics for this task focus on measuring how well Performer systems can generate definitions for novel specialized terms. System-generated definitions for specialized terms in the test set will be evaluated against hidden ground truth definitions. Systems will be evaluated on a combination of automated and human-derived metrics, which will be announced at program kickoff. The Government is especially interested in definition generation capabilities that work on multilingual collections and accurately generate likely definitions despite limited attestations and context.

As measure of baseline performance, the T&E team will use a state-of-the-art large language model prompted to generate definitions. Additionally, evaluation will look at how well systems perform on the definition generation task with progressively smaller amounts of available context.

Program Waypoints, Milestones, Deliverables

Decipher is a single phase 12-Month program with four evaluation milestones to evaluate Performers' progress to meet the program metrics. The program will culminate in the delivery of documented software that can be easily deployed by Government analysts using a standard laptop with a Graphical Processing Unit (GPU). Table 1 shows a timeline for the program with Government-defined milestones and deliverables.

DRAFT

Table 1: Program Waypoint, Milestone and Deliverables Testing Timeline

Event	Deliverables	
	Months after Kickoff	
Program Wide Milestone Kickoff	1	Read-ahead package due from Performers to the Government 7 days before meeting. If required by the PM, updates after the meeting are due 15 days after the meeting date.
Technical Review Meetings	Every 2 Weeks	Read-ahead package due from Performer to the Government 1 day before meeting. If required by the PM, updates after the meeting are due 7 days after the meeting date.
Site Visits and Waypoint Review	3,7	Site visits
Evaluation Milestones (Software and Data Deliverables)	2,5,8,11	Performers will submit their software, development datasets and documentation to the T&E partner on the first business day of the month. Source code, build scripts and documentation for facilitating reproduction of system operation by an independent Test and Evaluation Team. Performers shall also provide documentation and training to the Government for proper usage. All deliverables must be clearly labelled to identify IP assertions, which may be UNLIMITED or Government Purpose Rights.
Independent T&E	3,6,9,12	Upon receipt of the deliverables, T&E will be conducted. Performers may expect test results and analysis by the final business day of the month.
Financial and Technical Reports	Monthly	Monthly financial and technical reports are due by the 10th day of the following month in the IARPA provided template.
Delivery of deployable software with documentation	11	Performers will submit their software and documentation to the T&E partner on the first business day of the month. To enable transition of the capability, software should run on a standard laptop with a GPU. Performers shall also provide documentation and training to the Government for proper usage. The software packaging must clearly identify IP assertions, which may be UNLIMITED or Government Purpose Rights.
Collaborative Workshop	12	A Collaborative Workshop held in the DC area at the end of Program
End of Program, Final Report	12	Period of Performance Ends. Performers shall provide a final technical report outlining their technical progress, methodology, challenges, successes, and recommendations for future efforts.

Software Deliverable Formatting

Performers will be required to provide algorithms, software deliverables, design documentation, and operating instructions to execute the automated software in a manner that conforms to a standardized industrial method or methods that will be provided at program Kickoff. To facilitate planning, Offerors may assume that the standardized configuration will require the use of software containerization technology (e.g., Docker and a REST API). This means that the entirety of a Performer's system, including pre- and post-processing, must be included within the delivered software container. For models that require training, the expectation is for the initial model training to occur on Performer systems, with the ability for the T&E Team to re-train and test the model with the same and/or other data. Performers shall also provide source code to the government and T&E team to enable modification of Performer containers to suit deployment constraints. Offeror teams that do not include the requisite expertise to conduct such software development shall include costs in their white paper/proposal to obtain software development support.

Each team is required to include among their key personnel a Lead System Integrator (LSI) who shall be responsible for preparing software deliverable subcomponents, modules, and systems, performing quality control of Deliverables, and integrating key components into the primary Decipher system(s). The LSI will also oversee communication and coordination across Performer's research teams including subcontractors, if applicable, to ensure research products are functional and follow software coding best practices (e.g., inline comments, documentation).

Place of Performance

Performance will be conducted at the Performers' approved sites with Government's concurrence.

Test & Evaluation

T&E will be conducted by an independent team of Government and contractor staff carrying out evaluation and analyses of Performer Deliverables using program test datasets and protocols. Evaluations will occur during months 3, 6, 9, and 12 that will exercise Performer solutions across technical challenges.

Performers are encouraged to develop methodologies for a self-contained / self-hosted environment for use by government transition partners on their own network/systems. White papers must specify the processing dependencies needed to carry out the proposed research and what architecture and library characteristics are necessary for their approach(es) to be successful at meeting program objectives. Performers will have specific Deliverable Milestones at which all subcomponent and system algorithms and software will be delivered to IARPA and its designated T&E Team. The T&E Team will then conduct evaluations at the direction of the Decipher PM and with the objective of characterizing the quality, functionality, and performance of the Decipher Deliverables. In addition to quantitative measurements, T&E will be carried out to establish a thorough understanding of the progress, status, and limitations of the Performer's

DRAFT

research. T&E results and feedback will be provided to Performers at regular intervals to keep them abreast of current independent performance measurements and to inform and improve their R&D approaches and methods. T&E results from all Performers will be shared with all teams to establish an understanding of the current state and progress of Decipher research; T&E results will also be shared with United States Government external stakeholders, including their contractors, for Government purposes.

Technical Exchange Meetings/Workshops/Site Visits/Travel Requirements

Performers are expected to assume responsibility for administration of their projects and to comply with contractual and program requirements for reporting, attendance at program workshops, and availability for site visits. The following paragraphs describe typical expectations for meetings and travel for IARPA programs as well as the contemplated frequency and locations of such meetings. In addition to ensuring that all necessary details of developed software, algorithm, and operational instructions are clear and complete, each Performer will be required to be available for questions and troubleshooting from the T&E Team in weekly or bi-weekly status meetings.

All Performer teams are expected to attend a collaborative workshop, to include Key Personnel from prime and subcontractor organizations. The workshop will focus on technical aspects of the program and on facilitating open technical exchanges, interaction, and sharing among the various program participants. Program participants will be expected to present the technical status and progress of their efforts to other participants and invited guests.

Site visits by the Government Team will take place during months 3 and 7 during the life of the program. These visits will take place at the Performer's facility. Reports on technical progress, details of successes and issues, contributions to the program goals, and technology demonstrations will be expected during these visits.

Anticipated Period of Performance

Decipher is a 12-Month Program.

DRAFT