TAC 2023

CRUX: **C**laim **R**elation **U**nderstanding and **E**xtraction

Evaluation Plan

Last Updated: June 1, 2023

Revision History

V1: June 1, 2023

- Initial release

1 Introduction

In scenarios such as natural disasters or international conflicts, both professional analysts and the general public are often confronted with a vast amount of information coming through multiple media sources in multiple languages. In order to understand these situations and respond appropriately to them, there is a need for technologies that can analyze and extract claims from multilingual multimedia sources and determine how the claims are related.

The goal of the Claim Relation Understanding and Extraction (CRUX) track of NIST's Text Analysis Conference (TAC) is to evaluate technologies that extract explicit alternative claims from multilingual multimedia sources about situations and events in noisy, conflicting, and potentially deceptive information environments. The TAC 2023 CRUX track asks systems to extract claims that are made about the COVID 19 pandemic, including information about who is making the claim, their stance towards the claim, and when and where the claim was made. Systems are also asked to determine the relation between pairs of claims (e.g., whether one claim supports or refutes the other). The output of a CRUX system may be useful for downstream efforts to identify and characterize misinformation, disinformation, and influence campaigns.

This document describes the specifications of the tasks and evaluations for the TAC 2023 CRUX track. Although the track evaluates system performance using evaluation data developed under DARPA's Active Interpretation of Disparate Alternatives (AIDA) program, participation in the CRUX track is open to all researchers who find the evaluation tasks of interest.[1] There is no cost to participate. Participants are encouraged to attend a post-evaluation workshop at their own expense to present and discuss their systems and results. Information and updates about the tasks and evaluation will be posted to the CRUX website (https://tac.nist.gov/2023/KBP/CRUX/).

The CRUX 2023 track is conducting two tasks, in two phases:

- Phase 1: Claim Frame Extraction Task
    - Topic/data release for development: May 2023
    - Evaluation window: September 1-15, 2023
    - Evaluation annotation release: September 18, 2023
- Phase 2: Cross-Claim Relation Tagging Task
    - Topic/data release for development: May 2023
    - Evaluation window: September 18-30, 2023

[1] Participants in the CRUX track who were also performers in AIDA are already familiar with the CRUX tasks and topics and may have a potential advantage over non-AIDA participants; therefore, CRUX evaluation results will include an indication of whether the participant was in AIDA.

- Evaluation annotation release: October 7, 2023

The TAC 2023 Workshop will be co-located with the Text Retrieval Conference (TREC 2023) and will be held as a hybrid meeting on Nov 13-17, 2023 in Gaithersburg, Maryland, USA.

2. Track Overview

2.1 [Task I] Claim Frame Extraction

Definitions Following the DARPA AIDA program, we define each **claim frame** as a combination of values for the following fields of a claim frame template:

- *root_uid* - root UID associated with the source document from which the claim was extracted.
- *claim ID* - a unique ID for the claim frame.
- *topic (primary,* weight=0.19*)* - the topic of the current claim, selected from a predetermined list of topics of interest. Each topic has a corresponding ID and an English gloss.
  `Curing/Preventing/Destroying the Virus`
- *claim template (primary,* weight=0.19*)* - the unpopulated claim template for the current claim, selected from a predetermined list of claim templates for the topic. Each claim template is associated with one of the subtopics of the topic and has a corresponding ID and an English phrase.
  `X cures COVID-19`
- *X variable (primary,* weight=0.19*)* - a name or a short descriptive phrase that fills in the X variable to populate the claim template of the current claim.
  `lemon slices in hot water`
- *claimer (primary,* weight=0.19*)* - the entity (person, organization, etc.) making the claim.
  `Jiao Shenme Minzi`
- *epistemic status (secondary,* weight=0.16*)* - claimer's stance on the truth and certainty of the populated claim template.
  `{true-certain, true-uncertain, false-certain, false-uncertain, unknown}`
- *claimer affiliation (secondary,* weight=0.02*)* - affiliation of the claimer (possibly inferred); or EMPTY_NA if the information is unavailable in the source document.
  `Mayo Clinic`
- *sentiment_status (secondary,* weight=0.02*)* - claimer's sentiment toward the populated claim template.
  `{negative, positive, mixed, neutral-unknown}`

- *claim_datetime (secondary,* weight=0.02*)* - datetime type (before, after, on, unknown), and if known, the datetime (in yyyy-mm-ddTxx:xx:xx format) when the claim was made; or EMPTY_NA if no datetime is available in the source document.

  `on xxxx-04-04Txx:xx:xx`
- *claim_location (secondary,* weight=0.01*)* - location where the claim was made; or EMPTY_NA if the location is unavailable in the source document.

  `Loudoun County`
- *claim_medium (secondary,* weight=0.01*)* - the broadcast source (platform, channel, etc.) on which the claim was asserted; or EMPTY_NA if the information is unavailable in the source document.

  `CNN`

Each unique claim should be annotated exactly once per document (even if it occurs multiple times in the document), drawing on information from the document as a whole to complete the claim frame. A claim is considered unique if it differs from other claims in at least one of: topic, claim template, X variable, claimer, epistemic status, sentiment status, claim datetime, claim location, and claim medium.

Problem Statement Given a corpus of web documents and a list of targeted topics and per-subtopic claim templates, the goal is to extract a comprehensive set of unique claims about the targeted topics and subtopics from each document, along with their claim frame information as defined in the subsection above. (Example practice topics and subtopic/claim templates are provided in **Appendix A1**.) A per-document knowledge base (KB) of entities and keyphrases will also be provided for participants to fill in fields in the claim frame (e.g., X variable, claimer, claimer affiliation, etc.); this is further detailed in **Appendix A2**.

The **CRUX** claim frame extraction evaluation setting consists of 250 evaluation source documents related to the COVID-19 scenario, a list of 3 evaluation topics and their subtopics and claim templates, and a per-document KB of entities and keyphrases. Participating systems should extract all unique claims about the targeted topics and subtopics in the source documents and output a tab-separated file containing one line per claim frame.

An example line in the tab-separated *.tab* output file looks like this:

~~~
  ```
Doc ID B   Claim ID J      Curing/Preventing/Destroying the
Virus      X cures COVID-19     lemon slices in hot water
     Jiao Shenme Minzi    true-certain    Mayo Clinic
     neutral-unknown      on xxxx-04-04Txx:xx:xx    Loudoun
County     CNN
~~~

```
...
```

Scoring

We utilize the following metrics.

- Primary metric: Precision, Recall, F1 between system claim frames and ground truth claim frames, counting true positives (TP), false positives (FP), and false negatives (FN)
    - Procedurally, we align system-extracted claim frames with the ground truth claim frames using maximal bipartite matching, in which at most one ground truth claim frame ($\boldsymbol{g}$) will be matched with a system-extracted claim frame ($\boldsymbol{s}$), and vice versa. The matching heuristic will be based on scoring candidate matches ($\boldsymbol{M_{s,g}}$) between $\boldsymbol{s}$ and $\boldsymbol{g}$ via a weighting function, $\boldsymbol{w}()$:

      Eq 1 $$\boldsymbol{w}(\boldsymbol{M_{s,g}}) = \sum_{\boldsymbol{f} \in \boldsymbol{set\ of\ claim\ frame\ fields}} \boldsymbol{w_f}(\,\boldsymbol{I}(\boldsymbol{f_s}, \boldsymbol{f_g})\,)$$

      `if` *the corresponding values in the topic, claim template, and X variable fields all match*`; else`

      0

      We consider that each pair of claim frames in a candidate match consists of a set of field values (topic, claim template, etc.) to compare for correctness in extraction. We first determine whether the system generated field value ($\boldsymbol{f_s}$) and corresponding ground truth field value ($\boldsymbol{f_g}$) match for each field f, through an indicator function $\boldsymbol{I}()$ which assigns a score of 1 for a match in a given field and a score of 0 for a non-match. Then, we compute a weighted sum of the match scores between the field elements within a candidate match of claim frames. Sec. 2.1 defines the weight for each field of a claim frame.

    - After we obtain a maximal bipartite matching between system claim frames and ground truth claim frames using the weights defined in Eq 1, a system claim frame $\boldsymbol{s}$ is counted as a FP if *either* i) $\boldsymbol{s}$ is not aligned in the maximal bipartite matching *or* ii) $\boldsymbol{s}$ is aligned with a gold claim frame $\boldsymbol{g}$ but $\boldsymbol{w}(\boldsymbol{M_{s,g}})$=0. FN is the number of ground truth claim frames that are not aligned in the maximal bipartite matching, while TP is the number of pairs of claims that are aligned and have non-zero weights.

    - In addition to F-score, we also compute the overall accuracy of the claim frame field extractions based on the average of the scaled bipartite matching scores across the aligned claim frames.

- Secondary metric: separate P/R/F1 score for each field in the claim frame template:
    - For each field, we compute the set of unique values, S, returned by the system for that field (regardless of how many claim frames it appears in); similarly, we compute the set of unique values, GT, returned in the ground truth for that field; we then compute precision/recall/F1 of S and GT.

2.2 [Task II] Cross-Claim Relation Tagging

Definitions Claims across a corpus of documents can be characterized by underlying **cross-claim relations**.

- *Identical* - two claims are **identical** if all of the following are true:
    - Same topic, subtopic and claim template
    - Same X variable identity
    - Same claimer identity
    - Same epistemic status truth value
- *Refute* - Claim A **refutes** Claim B if all of the following are true:
    - They are on the same topic
    - They are not identical
    - If Claim A is true, Claim B cannot be true
- *Support* - Claim A **supports** Claim B if all of the following are true:
    - They are on the same topic, not necessarily the same sub-topic
    - They are not identical and not refuting
    - They can both be true at the same time
    - If Claim A is true, Claim B is more plausible

  Note: Claim A supports Claim B doesn't mean Claim B supports Claim A.

- *Related* - Claim A is **related** to Claim B if all of the following are true:
    - They are on the same topic
    - They are not identical, refuting, or supporting
    - If Claim A is true, it doesn't affect the plausibility of Claim B

Examples of cross-claim relations are included in **Appendix A3**.

It is worthwhile to point out that, in general, cross-claim relations can also be *unrelated*. However, in CRUX, we focus on predicting the relations only between claims about the same topic. This is because our novel concept of fine-grained claim frames, which involves analyzing claims from an information element perspective, follows specific predefined templates based on the topic. This makes it a trivial task to judge whether claims are unrelated simply through examining their template nature. By concentrating on claims within the same topic, we ensure a meaningful analysis of the relationships between claim frames.

Problem Statement Given a set of claims with their corresponding background documents that are released after Phase I Claim Frame Extraction, the goal is to identify all possible cross-claim relations following the Sec 2.2 definition.

Scoring Cross-claim relation tagging is a multi-class classification problem. We use macro-F_1 as the official evaluation measure, which accounts for class imbalance better than micro-F1. Macro-F1 calculates the 1-vs-rest F-score for each relation category (i.e., *identical*, *support*, *refute*, and *related*), and then takes their unweighted average.

Participants are expected to enumerate all possible pairs of claim frames having the same topic from the given input set of claim frames and generate a tab-separated *.tab* output file in the following format:

```
Claim ID 1     support   Claim ID 2
Claim ID 1     refute    Claim ID 4
…

```

2.3 Data Details

Source Corpus In the TAC CRUX track, we utilize web documents that are related to the COVID scenario from English, Spanish, or Russian media sources. These documents are multi-media in nature and may contain various modality types including text and images, though the majority of the claim-relevant information comes from the text within the documents.

Annotation The data statistics for the development and evaluation settings are summarized below. Information about the number of unique claims in the evaluation setting are withheld from participants until after all submissions are due.

	Dev	**Eval**
# of docs	637	250
# claims (by different claimer, etc.)	637	-
# claims (after coreference on the core semantics)	58	-

# topics	11	3

Appendix

A1. Example List of Predefined Claim Topic, Subtopic, and Claim Template

Topic	Subtopic	Claim Template
Contracting the virus	Who contracts the virus	X can catch COVID-19
Transmitting the virus	What transmits the virus	X transmits/transfers COVID-19
Curing/Preventing/Destroying the Virus	Destroying the virus	X destroys COVID-19

A2. Examples of Knowledge Base Elements given for X Variable, Claimer, Claimer Affiliation, Claim Location, and Claim Medium

The task organizers will prepare a curated, comprehensive list of extracted entities and keyphrases from each document. For Task 1 claim frame extraction, CRUX participants will refer to the appropriate entities or keyphrase mentions, by their Knowledge Base (KB) identity, when outputting *X Variable*, *Claimer*, *Claimer Affiliation*, *Claim Location*, and *Claim Medium* values. The purpose of this KB identity mapping is to make scoring discrete and straightforward.

KB Identity	Canonical Mention of Entity or Keyphrase	File ID
NILQE30318	Vitamin C	L0C0499BP
NILQE30319	Vitamin E	L0C0499BP
NILQE30349	Masks	L0C049JWW
NILQE30350	Handwashing	L0C049JWW
…	…	…

A3. Examples of Cross-Claim Relations

Identical:

	Claim A	Claim B
Natural Language Description	Donald Trump says that the US and Mexican governments agreed to ban recreational and tourist trips	Trump says that the U.S. made the correct decision to lock down
Topic	Government actions related to the virus	Government actions related to the virus
Subtopic	Population restrictions related to the virus	Population restrictions related to the virus
Claim Template	[Government-X enacted population restrictions related to COVID-19]	[Government-X enacted population restrictions related to COVID-19]
X Variable	US government (Q48525)	The U.S. (Q48525)
Claimer	Donald Trump (Q22686)	Donald Trump (Q22686)
Epistemic Status	true-certain	true-certain

Refute:

	Claim A	Claim B
Natural Language Description	Nothing cures COVID-19	Taking hot baths cures COVID-19
Topic	Curing the virus	Curing the virus
Subtopic	Curing the virus	Curing the virus
Claim Template	[X will cure covid-19]	[X will cure covid-19]

X Variable	Nothing (TBD)	Taking hot baths (TBD)
Claimer	Unknown (TBD)	Unknown (TBD)
Epistemic Status	true-certain	true-certain

Support:

	Claim A	**Claim B**
Natural Language Description	A theft from a Canadian lab is associated with the origin of COVID-19	The Canadian government created COVID-19
Topic	Origin of virus	Origin of virus
Subtopic	Events associated with the origin of the virus	Who created the virus
Claim Template	[event-X is associated/involved with the origin of COVID-19]	[X created SARS-CoV-2]
X Variable	A theft from a Canadian lab (TBD)	The Canadian government (Q422404)
Claimer	Unknown (TBD)	Unknown (TBD)
Epistemic Status	true-certain	true-certain

Related:

	Claim A	**Claim B**
Natural Language Description	COVID-19 can be transmitted in hot, humid conditions	Pets transmit COVID-19
Topic	Transmitting the virus	Transmitting the virus

Subtopic	How virus is transmitted	What transmits the virus
Claim Template	[COVID-19 can be transmitted in X conditions]	[X transmits COVID-19]
X Variable	Hot, humid conditions (TBD)	Pets (TBD)
Claimer	Unknown (TBD)	Unknown (TBD)
Epistemic Status	true-certain	true-certain