MultiVENT 2.0: A Massive Multilingual Benchmark for Event-Centric Video Retrieval

Supplementary Material

9. Query Creation - Annotation Guidelines

In this section, we provide the guidelines given to our professional linguists for the task of writing queries targeting aspects of new events found within our larger collection sampled from InternVid, described in Section 4. This annotation task was set up and run in Label Studio, a flexible data labeling platform that allows for multimodal annotations.[5] When writing queries for MULTIVENT-TRAIN and MULTIVENT-TEST, we ran two iterations of this task: the first looked at a random sample of 200 Arabic, Chinese, English, Korean, and Russian videos. For MULTIVENT-TEST we added Spanish as a sixth language. We focused on annotating InternVid videos from the News & Politics and Sports domains, as these were the categories with the most event-focused content.

After this initial round, there was not sufficient raw event-focused videos. To address this, we again sampled 200 videos per language from InternVid, this time filtering out any videos with standard resolutions, and continued until we had at least 20 annotated videos for each language/video type pair.

9.1. Task Introduction

The goal of this task is to annotate videos based on event templates to identify relevant information/entities across multiple modalities (text descriptions, video footage, text in videos, and audio/speech). These annotations were used to create queries based on the MULTIVENT 1.0 dataset and the events it contains. For the retrieval task to be meaningfully challenging, the set of documents from which systems are retrieving needs to be quite large (in the hundreds of thousands). As such, researchers have supplemented the MULTIVENT dataset with videos from a larger collection of videos, INTERNVID, and have begun writing queries associated with videos from that set.

In this annotation task, language experts will write queries for INTERNVID videos, with a focus on queries that come from specific modalities. For each annotation, you will be presented with a video (from YouTube) and its description (if present). Your task will be to watch the video, read the description, and write up to 4 English search queries based on information in the video, video description, video audio, and video text.

[5] https://labelstud.io

9.2. Video Relevance

What is an event? We broadly define an "event" as some set of properties and their changes (or lack of change) over time. A single, distinct event can only exist at one set location(s) across one set time interval. For this task, we generally constrain an event to something that occurs naturally, that is, not a staged set of predetermined actions (think movies, tutorial videos, etc.). We generally want to write queries that pertain to the main events depicted in the videos.

When should I skip a video? If the video falls into one of the below categories, select *Not Relevant* and submit the annotation.

- If it is hard to tell what the main event in the video is.
- If there is minimal event-relevant information present in the video (for example, a black screen, or a slideshow of irrelevant images).
- If the video is staged or otherwise not natural (movies, commercials, tutorial videos, etc.).
- If the video is not mostly in the language you are annotating for (this applies to the description as well as the audio content).
- If you are uncomfortable watching the video. Please do not annotate any videos that you don't want to watch!

9.3. Video Classification and Video Type

Select an event type that best describes the event in your video. If none of the listed event types apply, select *Other event type*.

- Disaster (fire, earthquake, hurricane, etc.)
- Political Election
- Political Protest
- Other Political Development (sanctions, treaty, leader death, etc.)
- Social Event (festival, convention, celebration, etc.)
- Sporting Event
- Technical/Scientific Launch or Discovery
- Other Event Type

Then, indicate the type of video:

- Professional news broadcast
- Non-professional edited footage
- Mostly raw footage
- Raw footage

The guidelines for video type classification are:

- The video is **professional footage** if you can imagine the clip appearing on broadcast TV. For example, there is an anchor/reporter, there are news logos and text banners with headlines, etc.
- The video is **non-professional edited footage** if it is not professional AND there is superimposed OCR/graphics, other visual special effects have been applied, or there are multiple clips spliced together. For example, many YouTube streamers' videos are "edited non-professional."
- The video is **mostly raw footage** if the video is a mostly continuous raw stream of content but has minor additional content (i.e., one of speech/music overlay, text overlay, or minimal scene cuts).
- The video is **true raw footage** if it is a continuous, unedited stream of footage.

9.4. Query Writing

Tips for good query writing:

- All of the queries should be in English.
- Write queries in the style you would use for a search engine, e.g., Google. The queries should be short and do not need to be written in complete sentences.
- Only write queries based on information that would help you learn more about the event in question. There may be video, audio, or text content that is not relevant to the main event topic—these irrelevant details should not be the basis for your queries.
- The queries should be unique from one another. If you cannot write a query for unique information in a particular modality, check the relevant box for why you left the query blank (see below).
- You can use supplemental internet searches to clarify information contained in the video description, e.g., to check spelling for a person's name.
- There are multiple "right answers" for each video's queries.

Each query has an accompanying text box; write your query and hit Enter to submit that query. Once submitted, the query will appear in a green text box.

9.4.1 Base Event Query

Watch the video. You may also read the title/description to get a better sense of what the video is depicting. Then, write one query based on the **primary topic/event** depicted in the video. This should include general information about what happened. The goal is to capture specific events. Each query should be approximately 3-7 words. An example of this question is provided in Figure 4.

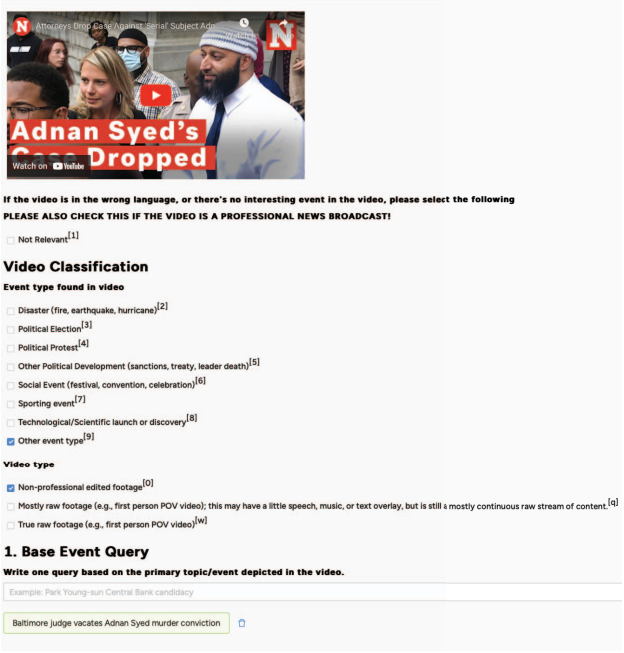


Figure 4. Example and Label Studio interface for video classification annotations and base query creation i.e., the phrase that represents the "Wikipedia title" of a current event.

9.4.2 Text-based Query

Write one query based on event-relevant details from the video description. This query should cover information that is not already addressed by previous queries. The query should be approximately 3-7 words. An example of this question and its associated interface is provided in Figure 5a.

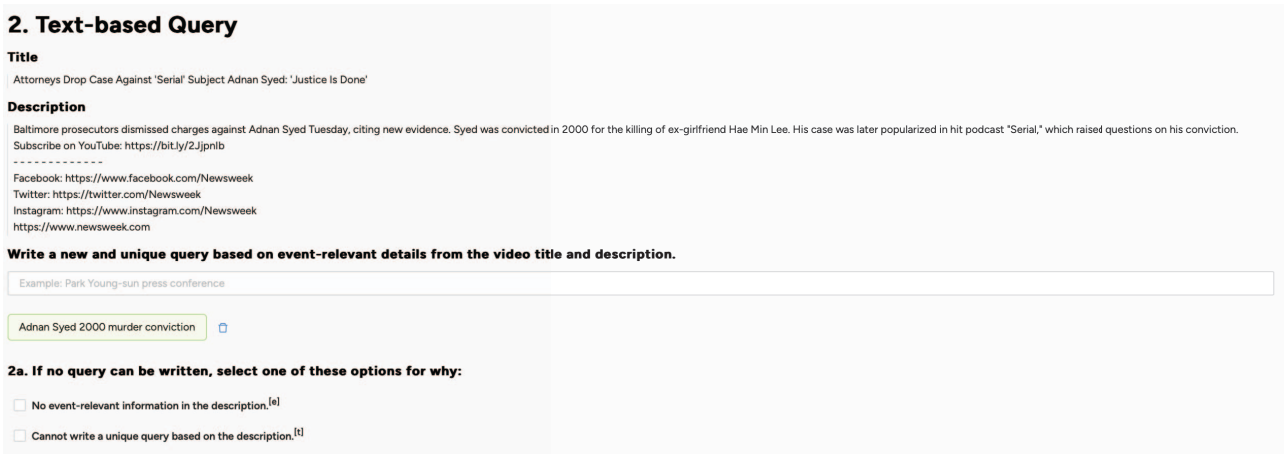
If no unique query can be written, select an option explaining why:

- No event-relevant information in the description.
- Cannot write a unique query based on the description.

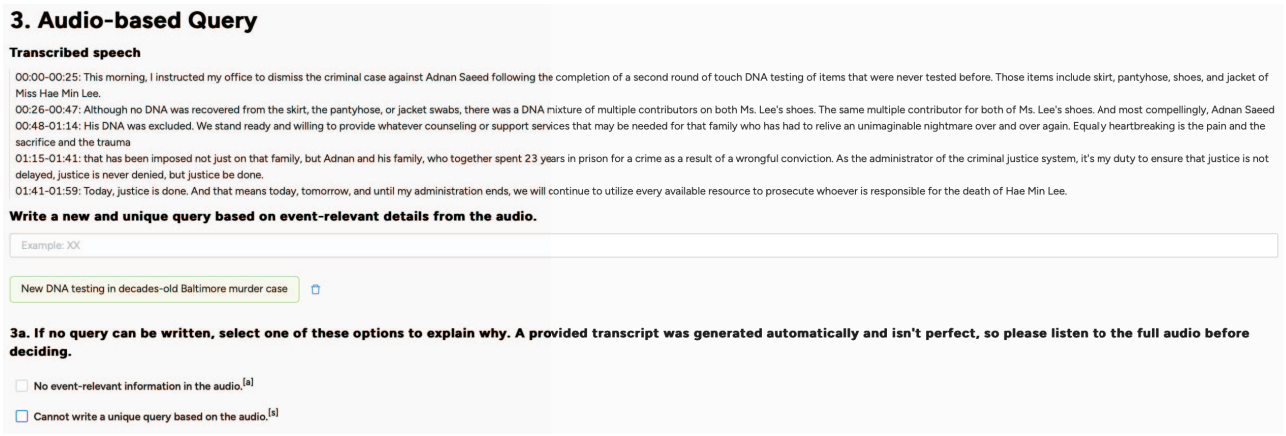
9.4.3 Audio-based Query

Write one query based on event-relevant details from the video audio/speech. This query should cover information that is not already addressed by previous queries. The query should be approximately 3-7 words.

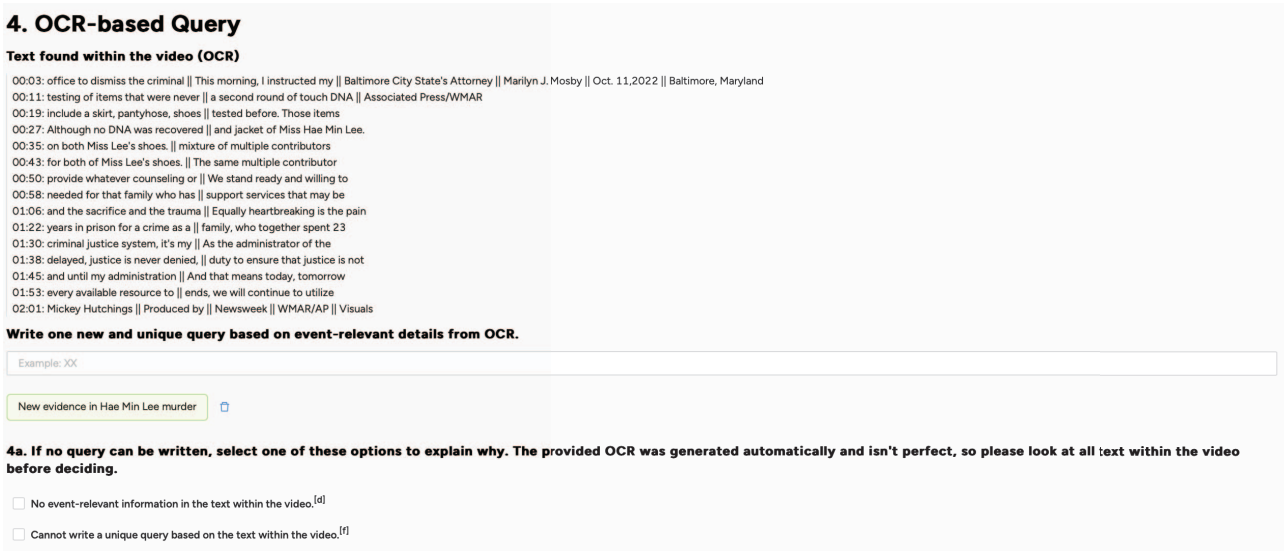
To aid annotation, transcribed speech has been provided (in the original language or translated, depending on the project you are working on). Note that these transcriptions, especially once translated, can have errors. Listen to the audio if you are a language expert and try to confirm spellings of names, places, etc. before finalizing your queries. An example of this question and its associated interface is provided in Figure 5b.



(a) Query creation using only the associated text description



(b) Query creation process using only audio, with associated transcribed text for assistance



(c) Query creation process using only embedded text, with associated OCR output for assistance

Figure 5. Example and Label Studio interface for writing queries targeting specific aspects of events. For each question, we ask annotators to only use the text description, audio, and embedded text, respectively.

If no unique query can be written, select an option explaining why:
- No event-relevant information in the audio.
- Cannot write a unique query based on the audio.

9.4.4 OCR-based query

Write one query based on event-relevant details from text in the video footage (e.g., signs, captions, etc.). Read the text found within the video (*OCR*) section. This query should cover information that is not already addressed by previous queries. The query should be approximately 3-7 words.

To aid annotation, OCR transcripts have been provided (in the original language or translated, depending on the project you are working on). Note that these transcriptions, especially once translated, can have errors and omissions. Consult the video footage for text that may have been missed or misrendered by OCR and consider taking a screenshot and re-translating in Google Translate if you are not a language expert before finalizing your queries. An example of this question and its associated interface is provided in Figure 5c.

If no unique query can be written, select an option explaining why:
- No event-relevant information in the text within the video.
- Cannot write a unique query based on the text within the video.

10. Relevance Judgment - Annotation Guidelines

In this section, we provide the guidelines given to our professional linguists to judge the relevance of previously unseen query/video pairs, described in Section 5.

10.1. Introduction

The goal of the task is for models to differentiate, i.e., through a ranked list, between relevant and irrelevant documents with respect to each query. Using Google as an example use case, a user expects the content most relevant to their search to appear in the first page of results. Through our prior query creation, we have developed a set of over 3,900 unique queries. To make the retrieval task meaningfully difficult, we have also included a large set of distractor videos (videos for which we have not written queries) in the overall document set.

When evaluating models on the retrieval task, our metrics assume that any given unseen video will be *Not Relevant* for any given query. However, it is possible that there are videos in our distractor set that are relevant for some of our queries. We want to find as many relevant videos as possible for a query to reduce the possibility of accidentally labeling something not relevant that actually is (called a *false negative*). This means potentially increasing the number of videos that are judged as relevant. This will help us more accurately judge systems.

Providing Relevance Judgment annotations involves assessing a query-video pair and determining if, for that query, the video is *Very Relevant*, *Somewhat Relevant*, or *Not Relevant*. The sections below provide more detailed guidance for this task.

10.2. Annotation Instructions

For each annotation task, you will be presented with a video (from YouTube), its title and description (if present), and extracted speech and embedded text. The text will be translated to English, if necessary (if you are a language expert and want to see the original title/description, open the video in a separate window in YouTube.

10.2.1 Video Type

After watching the video, indicate the type of video—professional footage/news broadcast, non-professional edited footage, or raw footage (i.e., first person POV video). An example of this component is shown in Figure 6a.

The guidelines for video type classification are:
- The video is **professional footage** if you can imagine the clip appearing on broadcast TV. For example, there is an anchor/reporter, there are news logos and text banners with headlines, etc.
- The video is **non-professional edited footage** if it is not professional AND there is superimposed OCR/graphics, other visual special effects have been applied, or there are multiple clips spliced together. For example, many YouTube streamers' videos are "edited non-professional."
- The video is **mostly raw footage** if the video is a mostly continuous raw stream of content but has minor additional content (i.e., one of speech/music overlay, text overlay, or minimal scene cuts).
- The video is **true raw footage** if it is a continuous, unedited stream of footage.

Note: some news agencies/platforms release a mix of video types. A "non-professional edited" video may be from a professional new company, but not meet our criteria for professional footage.

10.2.2 Relevance Judgments - Overview

Next, you will see a list of 1 or more English queries that you will judge that video against for relevance. For each query, you will select one of the following options:
- **Not Relevant**

- **Possibly Relevant**, i.e., the query mentions a clear event, and the video is possibly about that event
- **Partially Relevant**, i.e., the query mentions a clear event, the video is about that event, but does not show everything in the query
- **Very Relevant**

You will also have a text box to optionally leave comments (e.g., if you are not sure and/or escalating the judgment and want to briefly explain your reasoning).

10.2.3 About Our Queries

Base event queries When we wrote the base event queries, we tried to identify the main event being addressed by a given video. The model for this approach was MultiVENT base event queries, which were based on trending events that i.e. were covered by a Wikipedia article. This means that often the base event queries are broader than the videos they were written for, which may only cover specific aspects of the event in question. Some examples are below:

- A video covering loosening mask requirements in March 2022 in Canada might have the base event query *Covid-19 pandemic in Canada*.
- A video depicting a post-game interview with Chiefs quarterback Patrick Mahomes might have the base event query *Super Bowl LVIII*.
- A video of highlights from a presidential debate might have the base event query *2022 South Korean presidential election*.

For each of the query-video pairs above, the video would be Very Relevant for that query, even though it may not address the event as a whole. The base event queries may also name the event using details not explicitly mentioned in the video (for example, the year). However, if you are able to determine that the event described in the query is the same underlying event that the video is about, that video will be *Very Relevant*.

Specific Queries We wrote specific queries based on details found in videos. Specific queries will usually refer to the event in question as well as a specific aspect of that event. When you judge relevance for specific queries, you will want to consider (to the extent that they are decomposable) both the event described by the query and the specific aspect(s) or detail(s) about the event included in the query. Some examples are below:

- For a query *Black Sea drone incident damaged propeller*, you will want to consider 1) is the video about the Black Sea drone incident and, if yes, 2) does the video also include reference (in the image, description, speech, and/or embedded text) to a damaged propeller?
- For a query *Democratic Party response to Daejang-dong scandal*, you will want to consider 1) is the video about the Daejang-dong scandal, and 2) does the video also include content about the Democratic Party's response to the scandal?

If a specific query does not explicitly reference the *base event*, evaluate relevance based on each detail in the query. Some examples are below:

- For a query *Eric Adams hiring Director of Rat Control*, consider whether the video explicitly refers to Eric Adams (or New York Mayor) and to Director of Rat Control/Director of Rodent Mitigation.
  - If the video is about Eric Adams but unrelated to rat control, the video is Not Relevant.
  - If the video is about hiring the Director of Rat Control but does not refer to Eric Adams/the New York Mayor, the video is Partially Relevant.
- For a query *Canada eases entry restrictions September 2022*, consider whether the video is about updates to entry restrictions, in Canada, in September 2022.
  - Because the query points to a specific event, if the video is about Canadian entry restrictions not in September 2022, the video is Not Relevant.
  - If the video is about Canadian entry restrictions and the date/time period is not clear, the video is Partially Relevant.

10.2.4 Consulting Outside Sources

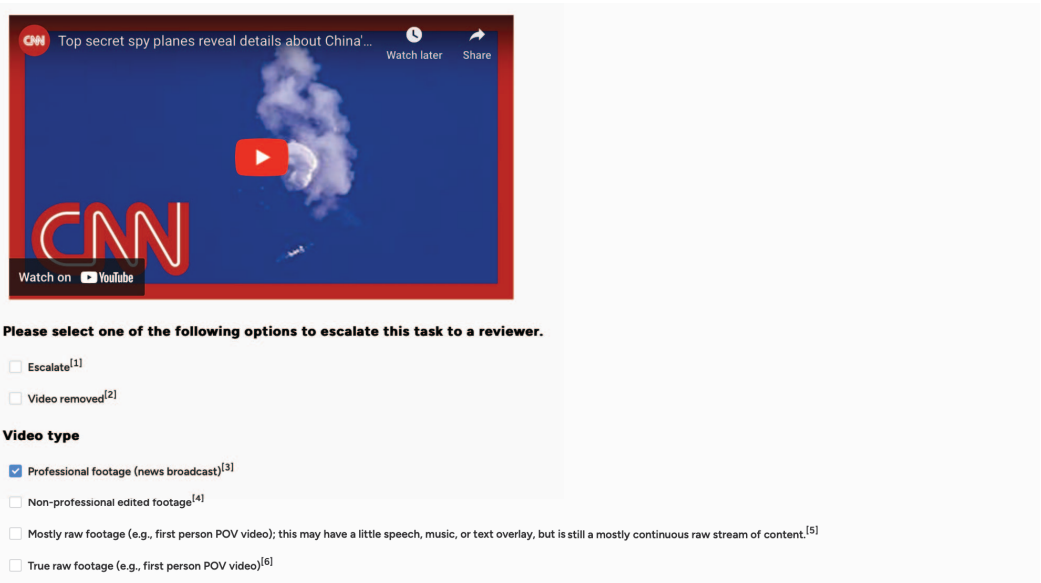
If you are unfamiliar with the event, people, places, etc. described by a given query, we encourage you to search for the event and/or entities online to learn more about the query's context. For named entities, this is particularly helpful, because the queries might refer to those entities in different ways than the videos do. For example, a query might say *Norilsk avalanche* and a video might refer to an *avalanche in the Krasnoyarsk region*; if you are not familiar with the region, an internet search may be required to determine that the video is relevant for that query.

When assessing the queries and videos for relevance, try to think conceptually about what a query is asking for, rather than looking for that exact wording in the video.

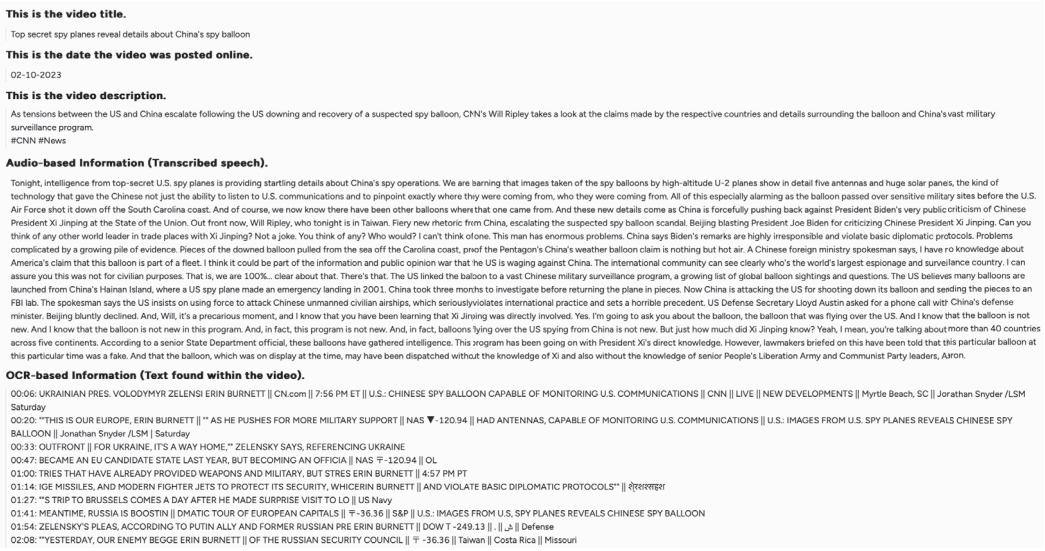
10.2.5 About Our Videos and Metadata

Most videos you will see for this task are from YouTube. You may also encounter some from Twitter. In addition to the video, you will also see various metadata for that video, shown in Figure 6b. For videos not originally in English, the metadata will have been translated into English.

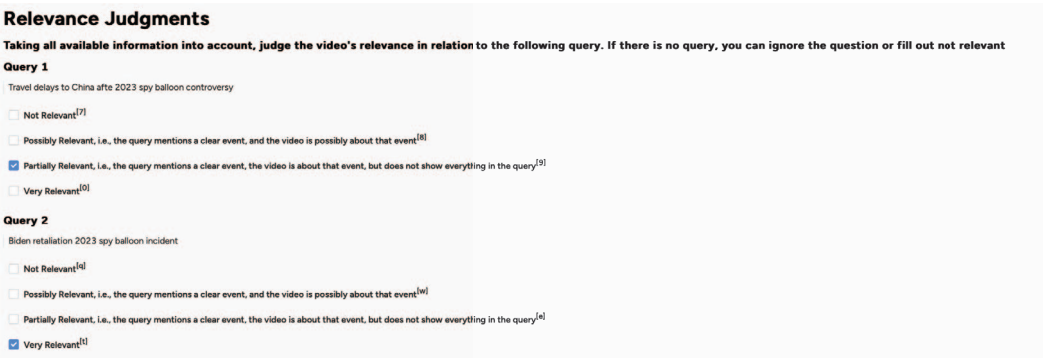
- **Video Title + Video Description.** If you are a language



(a) Example video and classification annotations



(b) Provided text extracted from the video



(c) Queries for which the video is judged for relevance.

Figure 6. Example of the annotation process for judging relevance for previously unseen query/video pairs.

expert and want to see the title and description in the original language, open the video in another window.
- **Posted Date.** Note that while it is helpful to use this date as a clue for the event depicted by the video, the date the video was posted online will not necessarily correspond closely to the date of the event. Consider the date along will all other information available to you.
- **ASR (automated speech recognition) output.** Speech from the video has been automatically transcribed, then translated into English if necessary.
- **OCR (optical character recognition) output.** A sample of text from the video (i.e., captions, banners, signs) has been automatically transcribed, then translated into English if necessary. Not all video frames were considered for this OCR output—this means that there may be text that was not caught at all for transcription and translation. You can supplement this information by pausing the video on a frame that has text on it, taking a screenshot of the video frame, and pasting the image into Google Translate for images.

Each automated process described above has the potential for errors. This is especially true for the recognition and translation of proper nouns, e.g., names. Consider all of the information available in the video, metadata, and outside sources to judge whether details in the video correspond to details from the query.

10.2.6 Relevance Judgments - Detailed Guidance

A video is Relevant with respect to query if some content in the video addresses the information-seeking need defined by the query. How primary that relevant content is within the video (e.g.,, if it is the main topic of the videos vs. If it is mentioned in passing) does not affect the video's relevance for the query.

Each query defines what kind of information we are looking for in the video (see discussion of Base Event queries and Specific queries above). Some videos will be Very Relevant (i.e., they completely address all aspects of a query), others will be Not Relevant (they do not address what the query is looking for). There are also two intermediate relevance judgments, Partially Relevant and Possibly Relevant. Example queries are provided in Figure 6c, and more details about each category are below:

Very Relevant [3] A video is *Very Relevant* for a query if it addresses all components of a query.
- For a base event query, a *Very Relevant* video is one about that event.
- For a specific query, a *Very Relevant* video is one that is about the event described by the query (if applicable) and includes the specific information/details in the query.

Partially Relevant [2] A video is *Partially Relevant* for a query if it addresses the general topic of the query but does not include the specific detail(s) the query is asking for.
- For a specific event query, a *Partially Relevant* video is one that covers the main event/topic in the query but is missing a specific detail included in the query.
- If you can confirm that a query is about an event distinct from the one in the query, the video is Not Relevant.

Possibly Relevant [1] A video is *Possibly Relevant* if it is unclear that the video is relevant for the query.
- For a base or specific event query, a *Possibly Relevant* video is one that may cover an aspect of the event, but it is not entirely clear that it does so (i.e., it is possible that the video is about the event/topic, but it could also be about a different event.).
- If you can confirm that a query is about an event distinct from the one in the query, the video is *Not Relevant*.

Not Relevant [0] A video is *Not Relevant* if it does not address the event described in the query. This includes videos that are about similar but distinct events—for example, a video about the *2018 Winter Olympics* when the query is about the *2022 Winter Olympics*.

When to escalate Escalate in the following scenarios:
- You are unable to access or play a video (before you escalate for this reason, try to open the video in another window). [Select *Video Removed*]
- The video is in the wrong language (in relation to the project you are working on)
- Given your understanding of the task and these guidelines, you are unable to determine the video's relevance in relation to one or more queries. In this scenario, please also post a message to the Teams chat describing your question. If your question is answered, you can return to the task later, update judgments as needed, and un-select the Escalate option. [Select *Escalate*]

Note that you can still provide relevance judgments for some queries, even if you escalate the task as a whole. If there are queries for which you can confidently judge relevance, please do so. If you escalate a task, leave a brief explanation in the comment box to describe your reason for escalation and any other details you think would be helpful for the next person evaluating the video.