MED + MER 2014 Evaluation FAQ

Updated 2014-07-16

1. **How do teams interface with the I/O Server? How does the NIST I/O Server call our modules? How does the I/O Server collect timing information?**

   See section 1 of the Evaluation Plan.

2. **What happens if the edited semantic query is NOT syntactically or semantically correct?**

   The semantic query should be clear enough for the judge so that they will not make syntactic or semantic mistakes in removing nodes. However, if issues do arise, NIST will work with the participants on a case-by-case basis.

3. **Are teams participating in MER required to generate recountings for all videos or just the MED identified positives?**

   Teams participating in MER should only generate recountings for MED-identified positives as determined by the rank threshold which maximizes MRo.

4. **Are recountings required to have detector results for every element of the query?**

   No, the recountings only show evidence that was used to compute the event confidence score. If the requested evidence was not found/used, the system should set the confidence score for that evidence to zero.

5. **When will the I/O Server interface, and xml schemas be released?**

   Initial XML schemas will be released by May 1st 2014. The release date for the I/O Server interface is May 8th, 2014.

6. **When is external training data allowed? Are external resources available to the user during Semantic Query Generation?**

   External training data can only be used to implement the metadata generation module. The Semantic Query must only contain tags that are used in Metadata Generation.

7. **Are the Semantic/Event Queries executed by NIST?**

   No NIST will not run Queries, the teams' Event Search modules is responsible for executing the queries.

8. **The "private" node in an Event Query could be very large if SVM or neural net parameters are included.**

   This node can include the actual parameters or a reference to a local data store. It is up to the Event Search module to de-reference resources through the private section.

9. **Can teams resubmit results that were previously submitted?**

   The phases of the evaluation plan were edited to include optional runs of the modules where resubmissions are allowed. If there are extenuating circumstances, NIST will decide about resubmission on a case-by-case basis.

10. **What functions are teams allowed to use in the Semantic Query to combine key evidence scores?**

    Teams are allowed to use any functions as long as they are understandable to a non-technical user. Teams just need to make what the system is doing obvious to the user in the semantic query.

11. **How should teams report ties in video rank for detection?**

    See the detection file description in section 1.4 of the Evaluation Plan.

12. **Can the Event Query Generation Module modify the Semantic Query, for example by identifying additional key tags, and adjusting the importance of tags?**

    Yes, tuning the semantic query based on exemplars is allowed. Setting weights is an excellent example of how this may be better for a system to compute rather than a user to have to guess. Of course, this only applies to the EQ-000Ex, EQ-010Ex, and EQ-100Ex queries and they many only use the input exemplars to modify the Semantic Query portion of the Event Query.

13. **It's specified that MG can be run on any computing system, does this apply to all calls to MG, including the calls in PS – Part A (Sec 4.6) and Ad-Hoc (Sec 4.8)?**

    Yes, this applies to all calls to the MG module.

14. **Will the NIST I/O server be making the calls for each event independently or once for all events? Does this differ between modules?**

    Every call to a module should be run independently. A run of a module will process the inputs to produce its outputs for a single set of inputs. The inputs

and outputs of the modules are shown in the block diagrams for each module in Section 1 of the Evaluation Plan. A phase of the evaluation may have many runs of modules, but each module will be run independently.

15. Is it permissible to load all of the metadata for the Event Search (ES) at one time?

The Metadata for the Event Search should be loaded each time the Event Search module is called.

16. The annotations on the MEDTest for Events 31-40 are not yet available, who will provide it and when?

NIST will release the MEDTest annotations for Events 31-40 starting on April 28, 2014. The annotations on MEDTest for Events 31-40 will be provided via CD/DVDs mailed to teams, as soon as participants sign the licensing agreement.

17. How many system variants can we test for e.g. the 0x and 10x AH condition?

MED14 (as did MED13) only supports submission of results of one system per team. However, MED also provides a defined MED14Test set, which, if the team keeps this set separate from training activities, can be used for publication of system variants.

18. Last year, there was a proposal to ask teams to submit a run of their system on the MEDTest set and that this run would be evaluated by NIST and detailed results would be distributed to all the participating teams. Will this be supported?

We have no plan to do this at this time, but will consider it for MED15. Teams receive the ground truth reference data for MED14Test so teams are free to share results amongst themselves.

19. It is not clear to us what the research question is behind the semantic query editing pilot.

The research question behind the Semantic Query Editing Pilot is: "If a semantic query concise and clear, then a user can make simple edits to the query to form a new query for a new event". This year we are testing a reduced version of this (a more specific event) as a feasibility study.

20. Is the MED14Test set new each year (with a new set of events)? How many events are in the MED14Test?

MED14Test (and MEDTest last year) is provided as a development test set for internally evaluating algorithms and techniques. The background clip collection is held constant, but each year we add positives for the previous year's Ad-Hoc events. MED14Test will contain the 20 pre-specified events (E021-E040).

21. What is the function of the Metadata Store Description?

The file describes key aspects of the metadata generation process, resources, and components but not the metadata itself. It also tracks the files represented in the metadata store which we use for validation.

22. How can teams specify the relative importance of concepts in their queries?

Teams are allowed to include extra attributes in their queries and recounting outside of what is explicitly defined in the schemas. This allows teams to build queries/recounting that are more easily parsed by their system. The example shown below demonstrates how extra attributes can be used to specify the relative importance of concepts.

```
<node id="S1" name="Board Trick Object" eq="WEIGHTED_SUM">
        <tag id="S1.1" name="surfboard" weight="0.4"/>
        <tag id="S1.2" name="skateboard" weight="0.3"/>
        <tag id="S1.3" name="snowboard" weight="0.3"/>
</node>
```

Alternatively a node's equation ('eq') attribute can be used to specify not only the function used to combine scores, but also parameters and weights if necessary. The following are some notional examples of how teams can specify more complex equations using just the equation attribute ("ids" borrowed from the example query in Appendix A of the evaluation plan).

All subordinate 'score' attributes implied:

```
eq='MAX'
eq='SUM'
```

Explicitly specify subordinates referenced by node/tag id:

```
eq='MAX("S1", "S2")'
eq='SUM("S1", "S2")'
eq='"S1" + "S2"'
```

A weighted sum:

```
eq='SUM("D" => 0.4, "S" => 0.6)'
eq='("D" * 0.4) + ("S" * 0.6)'
```

In the context of a complete node element:

<node id='E001' name='Board Trick' eq='SUM("D" => 0.4, "S" => 0.6)'>
...
</node>

23. Is there any community effort to provide ASR transcripts for the MED HAVIC data? (i.e. making shared transcript available)

There is no organized community effort but it would be allowed based on the evaluation plan so long as the processing steps and reporting requirements are followed. For example: if TeamA shares ASR with TeamB:

- TeamA and TeamB need to check out the MG task at the same time (so that the start time is established for both teams).
- TeamA will provided content for TeamB's Metadata Info and Hardware Description files.

24. What does MED14-EvalFull consist of?

MED14-EvalFull will consist of two LDC data sets,

- LDC2012E26 - MED Video Data : Progress Test Collection (previously released)
- LDC2014E42 - MED-14 Eval Video Data: Novel 1 Test Data (to be released).

25. What does MED14-EvalSub consist of?

A subset of both LDC2012E26 and LDC2014E42.

26. Are Event-BG negative examples for events?

Event-BG is a random collection of videos to be used as a background video population. While there may or may not be similar videos to the event, Event-BG videos can not be annotated or changed in any way (Eval Plan Section 1.1).

27. Are teams allowed to use external resources (such as Wikipedia) tofind concepts related to those in their Metadata Store during SQG?

Yes, you may integrate these internet resources as part of your Event Query Language (a local resource) that is used as part of the Semantic Query Generation (SQG) module (see the SQG block diagram in the Evaluation Plan pg. 6).

28. How can the MED14Test data be used? Can it be used for model validation and tuning?

MED14Test is an internal testing collection meaning it should be used to evaluate the developed algorithms (just like the MED14Eval is used for the evaluations). Any cross-validation done during EQG must be done solely on the training videos provided as input to the EQG module.

29. Can we use near-miss videos in our training process, e.g. as positive or negative samples?

Yes, but you are restricted to the set of videos provided during the EQG execution.

30. What is the difference between 'DetectionThresholdScore' and 'DetectionThresholdRank'? How will they be used in the evaluation?

They measure the same thing but on two different scales: detection score vs. rank. During scoring, we will use the DetectionThresholdRank for computing R_o.

31. Will MAP be measured at a specific depth in the rank list?

For MED, MAP will be measured over all positive event instances.

32. For EK10, are we allowed to use extra/external concept detectors which are automatically inferred from the ten positive training videos?

No, you cannot use external concept detectors inferred from the training examples. However, you can infer any of the metadata tags currently in the metadata store (which use external concept detectors).

33. Can teams use the entire PSTraining set in training an event?

No teams can only use the positive and negative videos specified in the <eventID>_<ex>.csv input file. No other videos are allowed in training the event query.

34. Can teams use the complete set of PS event kit descriptions as a resource to compute video collection statistics for EQG? For MG?

MG and EQG must use only the information input to the module via the I/O server. No other information is allowed to be used. Since the complete set of PS event kits are not part of the input, they can not be used. However, teams may use the PS event kit text in selecting what concepts their system will tag in the metadata. These concepts must be trained and integrated into the MG before running the MG on any of the test data.

35. Which resources may be uses for deriving concept importance weights? Can we use collection statistics computed on the MED14 TEST or MED research or BG video collections?

Only the Research Set may be used for training concepts and assigning importance weights. All other data sets are used for testing and evaluation purposes and therefore should not be used to train concepts for MG.

36. Are teams allowed to modify the contents/structure of query elements inside of their recountings?

No, the recounting should only fill out the evidence for the event query, not change the event query in any way.