Evaluation Methods and Procedures for 2010 i2b2/VA Challenge

Evaluation Procedures

- Evaluation in three steps:
  - one for each of the tasks of the challenge
- Teams can submit up to three system runs per task
  - Each system run is a zip archive of one output file per input file
    - A concepts run consists of a zip archive of .con files
    - An assertions run consists of a zip archive of .ast files
    - A relations run consists of a zip archive of .rel files
- We will rank:
  - Systems that use public resources among themselves
  - Systems that include private or proprietary resources among themselves

Resources

- Public resources:
  - resources that are publicly available and equally accessible to all teams
  - To ensure equal accessibility, we ask the participants to share the names and references to their resources through the google group and post these references online up to a month before the evaluation time (deadline of June 22nd). This way, we will have a list of all resources that were accessible to all teams.
  - E.g., UMLS filtered based on some standards
- private or proprietary resources:
  - resources that are proprietary and/or that were not posted on the google group for the challenge
  - E.g., UMLS augmented with some data that is private

Data

- Target number for training records
  - ~450
    - 97 released on April 22
    - Rest will follow in installments. Next installment date is May 22
- Target number for test records
  - ~300
    - To be released on July 22
- The split into training and test is random
- Please note that the exact number of training and test records will depend on the number of annotations per record. We may release fewer records (that are richer in terms of annotations per record) or more records (that are poorer in terms of annotations per record) than the announced target numbers.

Evaluation Timeline and Procedure

We start at 22 July and go in 24 hour increments:

- Test data records release: 22 July (9am ET)
- + 24 hours:
  - System outputs on the concept extraction due
  - Ground truth on concept extraction release
- + 24 hours:
  - System outputs on assertions due
  - Ground truth for concept extraction can be used as input
  - Ground truth on assertions release
- + 24 hours:
  - System outputs on relations due
  - Ground truth for concept extraction and assertion classification can be used as input
  - Ground truth on relations release

Concept extraction task

- Three types of concepts: Diseases, treatments, and tests
- System input: raw text of medical records
  - Random split into different institutions and document types
- System output: A plain text file that contains entries of the form:
  - c="concept text" offset || t="concept type"
  - c="prostate cancer" 5:7 5:8 || t="problem"
  - c="chemotherapy" 5:4 5:4 || t="treatment"
  - c="chest x-ray" 6:12 6:13 || t="test"

Evaluation (task nature: Extraction)

- Primary metric: Exact Micro-averaged Precision, Recall, F-measure for all concepts together
- Exact Micro-averaged Precision, Recall, F-measure for diseases, treatments, and tests separately
  - Exact: Phrase boundaries and concept types match exactly
    - Correct boundaries with incorrect type get no credit
    - Incorrect boundaries with correct type get no credit
    - Incorrect boundaries with incorrect type get no credit
- Inexact Micro-averaged Precision, Recall, F-measure for all concepts together
- Inexact Micro-averaged Precision, Recall, and F-measure for diseases, treatments, and tests separately
  - Inexact: Concept tagged overlaps with the ground truth concept at least in part.
    - One partial match contributes a count of 1 to overall score
    - The total number of phrases in the corpus is the same as the number used for exact match

Assertion classification task

- Give types of assertions of problems: present, absent, possible, conditional, hypothetical or associated with someone else
- System input: raw text of medical records (same set as the concept task) and ground truth on concept extraction
- System output:
  - Assertions on all problem concepts (and only problem concepts)
  - Systems provide assertions on problem concepts (from the ground truth of the concept task), E.g., system output in a plain text file:
    - c="concept text" offset || t="concept type" || a="assertion value"
    - c="prostate cancer" 5:7 5:8 || t="problem" || a="present"
    - c="hypertension" 5:4 5:4 || t="problem" || a="absent"
    - c="diabetes" 6:12 6:12 || t="problem" || a="possible"

Evaluation (task nature: classification)

- Primary metric: Micro-averaged Precision, Recall, F-measure for all assertion types together
- Micro-averaged Precision, Recall, F-measure for each of the assertion types separately

Relation classification task

- Name the relation that holds between two concepts
- System input: raw text files (same files as before), ground truth for concepts (necessary), and assertions (optional)
- System output: relations of pairs of concepts in the following format:
  - c="a cardiac catheterization" 9:12 9:14 || r="TeCP" || c="chest pain" 9:5 9:6
  - c="a cardiac catheterization" 9:12 9:14 || r="TeRP" || c="an occluded right coronary artery" 9:23 9:27
  - c="a cardiac catheterization" 9:12 9:14 || r="TeRP" || c="a 40-50% proximal stenosis" 9:29 9:32

Evaluation (task nature: classification)

- Primary metric: Micro-averaged Precision, Recall, F-measure for all relation types together
- Micro-averaged Precision, Recall, F-measure for each of the relation types separately (and over groups of relations that are complementary.)