TAC 2018
Streaming Multimedia KBP Pilot

Hoa Trang Dang

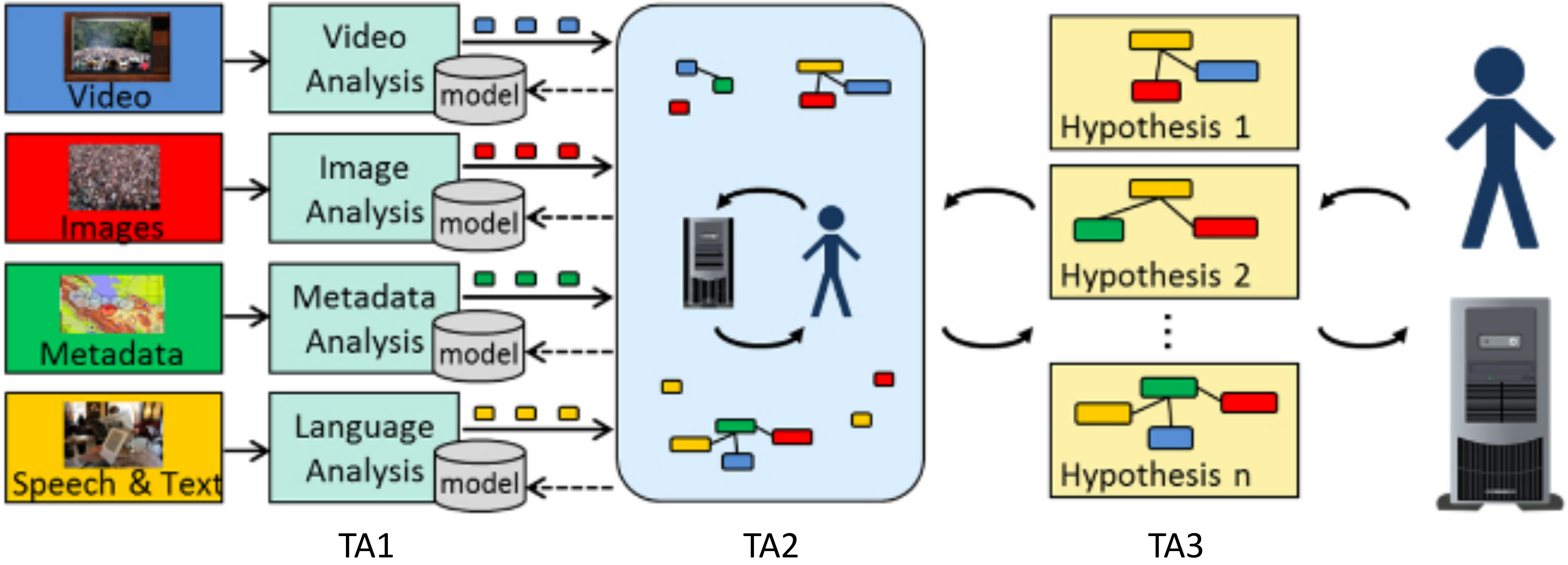
National Institute of Standards and Technology

Background

- NIST will evaluate performers in DARPA AIDA Program (Active Interpretation of Disparate Alternatives)
- Some AIDA evaluations will be open evaluations in TAC and TRECVID.
- The goal of AIDA is to develop a semantic engine that automatically generates multiple alternative analytic interpretations of a situation, based on a variety of unstructured sources that may be noisy, conflicting, or deceptive.
- Documents can contain a mix of multilingual text, speech, image, video; including metadata.
    - A document can be as small as a single tweet, or as large as a Web page containing a news article with text, pictures and video clips.
- All data will be in streaming mode; systems can access the data only once in raw format, but may access a KB containing a structured semantic representation of all data seen to date

ACTIVE INTERPRETATION OF DISPARATE ALTERNATIVES (AIDA)

- Given a scenario (“Benghazi”), document stream, and several topics. For each topic:
  - TA1 outputs all Knowledge Elements (entities, relations, events, etc., defined in the ontology) in the documents, including alternative interpretations
  - TA2 fuses KEs from TA1 into the TA2 KB, maintaining alternative interpretations
  - TA3 constructs internally consistent hypotheses (partial KBs) from TA2 KB



Scenario-Specific Ontology

- Scenarios will involve events such as international conflicts, natural disasters, violence at international events, or protests and demonstrations.
- AIDA will extend KBP ontology of entities, relations, events, belief and sentiment to include additional concepts that are needed to cover informational conflicts in each topic in the scenario
    - Ideally, would have a single ontology for all topics in the scenario (?)

AIDA KB representation

- Knowledge Element (KE) is a structured representation of entities, relations, events, etc. -- likely an augmented triple like in Cold Start KB
- Triple is augmented with provenance and confidence
  - Provenance is a set of justifications. Each justification has a justification-level confidence
  - KE-level confidence is explicitly provided by TA1 and TA2, and is an aggregation of justification-level confidences
- KB contains conflicting KEs (as found in the raw documents)
  - Representation -- not reconciliation -- of conflicts

What is allowed in KB representation?

- AIDA: “Although there may be need for some natural language, image thumbnails, featurized media, etc. in the KB for reference, registration, or matching purposes, it is expected that most of the assertions in the KB will be expressible in the structured representation, with elements derived from an ontology.”
- Features accessible to TA1/TA2 in KE cannot be document-level content features (?). Allowable features include
  - Number of supporting docs, and link to docs (but can’t read docs)
  - Time of first supporting doc, most recent supporting doc
- Comments/recommendations from participating teams are welcome regarding what features should be allowed in the KB
- For evaluation purposes, provenance accessible to LDC should be pointers into the raw documents denoting text spans, audio spans, images, or video shots

TAC/TRECVID 2018 tasks (pilot)

- Task 1: Extract all events, subevent or actions, entities, relations, locations, time, and sentiment from **multimedia** document **stream** , conditioned on zero or more different contexts, or **hypotheses** (TAC, TRECVID 2018)
    - Output is a set of all possible KEs, including confidence and provenance
    - Mention-level output, including within-document linking
- Task 2: Build KB by aggregating all KEs from TA1 and “user” (TAC2018)
    - Output is KB including cross-doc linking
    - Evaluate by queries (with entry points) and assessment
- [Task 3: Create hypotheses from Task 2 KBs (AIDA program-internal in 2018)]

Training/Evaluation data

- One new scenario per evaluation cycle; 4 scenarios total over lifetime of AIDA program.
- 100K docs/scenario, including relevant and irrelevant documents
  - 5-20% of docs will be relevant to the scenario
  - 200 labeled docs per scenario
- 12-20 topics per scenario
- At least one foreign language per scenario, plus English
  - AIDA: "Government will provide linguistic resources and tools of a quality and composition to be determined, but consisting at least of the type and size found in a LORELEI Related Language Pack (LRLP)"

Low Resource Language Packs

- 1Mw - 2Mw+ mono text from news, web text & social media
- 300Kw - 1.1Mw+ parallel text of variable quality (professional, crowd, found, comparable)
- Annotations for 25Kw - 75Kw/language including
  - Simple Named Entity (PER, ORG, GPE, LOC/FAC)
  - KB linking of names to GeoNames and CIA World Fact Book
  - Situation Frames: needs/issues for an incident (e.g. Urgent shelter need in Kermanshah province)
  - Full Entity (name, nom, pro) and within-doc coref
  - Predicate-argument annotation of disaster-relevant Acts and States
- Grammatical resources ranging from full grammatical sketch to found resources (dictionaries, grammars, primers, gazetteers) to lexicons
- Basic NLP tools including word, sentence segmenters, encoding converters; name taggers

Related TRECVID Tasks

TRECVID (2001 – Present)

- Shot boundary detection: Identify the shot boundaries in the given video clip(s)
- High-level feature extraction / Semantic Indexing: Given a standard set of shot boundaries and a list of feature (concepts) definitions, return a ranked list of shots according to the highest possibility of detecting the presence of each feature
- Ad-hoc Video Search: Given a statement of information need, return a ranked list of shots which best satisfy the need; similar to semantic indexing, but with complex concepts (combination of concepts); e.g., find group of children playing frisbee in a park.
- Rushes Summarization: Given a video from the rushes test collection, automatically create an MPEG-1 summary clip less than or equal to a maximum duration that shows the main objects and events in the rushes video to be summarized
- Surveillance event detection: detect a set of predefined events and identify their occurrences temporally
- Content-based copy detection: given a test collection of videos and a set of (video, audio, video+audio) queries, determine for each query the place, if any, that some part of the query occurs, with possible transformations, in the test collection

TRECVID (2001 – Present)

- Known-item Search: Given a text-only description of the video desired and a test collection of video with associated metadata, automatically return a list of up to 100 video IDs ranked by probability to be the one sought
- Instance Search: Given a collection of test videos, a master shot reference, and a collection of queries that delimit a person, object, or place entity in some example video, locate for each query the 1000 shots most likely to contain a recognizable instance of the entity [AIDA TA2 cross-doc coref]
- Multimedia Event Detection: Given a collection of test videos and a list of test events, indicate whether each of the test events is present anywhere in each of the test videos and give the strength of evidence for each such judgment
- Localization: Given a video shot, Determine the presence of a concept temporally within the shot, with respect to a subset of the frames comprised by the shot, and, spatially, for each such frame that contains the concept, to a bounding rectangle [AIDA provenance?]

Latest task introduced in 2016 : Video-to-Text

- Given a set of 2000 URLs of Twitter (Vine) videos and sets of text descriptions (each composed of 2000 sentences), systems are asked to work and submit results for two subtasks:

  - **Matching and Ranking:** Return for each video URL a ranked list of the most likely text description that correspond (was annotated) to the video from each of the different text description sets.

  - **Description Generation:** Automatically generate for each video URL a text description (1 sentence) independently and without taking into consideration the existence of text description sets.

- **Systems and annotators were encouraged to describe videos using 4 facets:**
  - **Who** is the video describing such as concrete objects and beings (kinds of persons, animals, things)
  - **What** are the objects and beings doing? (generic actions, conditions/state or events)
  - **Where** such as locale, site, place, geographic, architectural (kind of place, geographic or architectural)
  - **When** such as time of day, season, etc

Examples of concepts used in the TRECVID Semantic INdexing (SIN) task

Airplane
Anchorperson
Animal
Basketball
Beach
Bicycling
Boat_Ship
Boy
Bridges
Bus
Car_Racing
Chair
Cheering
Classroom
Computers
Dancing
Demonstration_Or_Protest
Greeting
Hand
Highway

Sitting_Down
Stadium
Swimming
Telephones
Throwing
Baby
Door_Opening
Fields
Flags
Forest
George_Bush
Hill
Lakes
Military_Airplane
Explosion_Fire
Female-Human-Face-Closeup
Flowers
Girl
Government-Leader
Instrumental_Musician

Oceans
Quadruped
Skating
Skier
Soldiers
Studio_With_Anchorperson
Traffic
Kitchen
Meeting
Motorcycle
News_Studio
Nighttime
Office
Old_People
People_Marching
Press_Conference
Reporters
Roadway_Junction
Running
Singing

Multimedia

- Each document can contain a mix of text, speech, image, video; including metadata.
- Multiple languages: English plus 1-2 foreign languages (TBA)
  - LDC will provide language packs containing resources for each language
- All participants will be given the same documents
  - Participants are allowed to process info in a proper subset of the languages or media types
  - NIST may report breakdown evaluation results by language, media type, etc.

Streaming Extraction

- Documents arrive in batches as a chunk.
  - ~100 documents/chunk (?), with cap on length of time covered in a chunk
- TA1 (and TA2?) system emits KE's (triple+confidence+extras) after each chunk.
- At specified time points in the stream, the set of accumulated KE's is evaluated.
  - Ranked precision/recall derivatives.
- At some of those points, a wild hypothesis appears!

  - A hypothesis = a set of proposed tuples.
  - TA1 system outputs KE's primed by the hypothesis, which are evaluated.

TA1 Extraction Conditioned on Context

- TA1 must be capable of accepting **alternate contexts** and producing **alternate analyses** for each context.
  - For example, the analysis of a certain image produces knowledge elements representing a **bus on a road**. However, knowledge elements in one or more hypotheses suggest that this is a **river rather than a road**. The analysis algorithm should use this information for additional analysis of the image with priors favoring a **boat**.
- Simplifying assumptions for evaluation purposes:
  - Contexts are coherent hypotheses (represented as a partial KB) drawn from a small static set of possible hypotheses that are produced manually by LDC
  - Only "what if" hypotheses are input to TA1; KEs and confidence values resulting from "what if" hypotheses do not get passed on to TA2 but are evaluated separately

How is Task 1 different from past TRECVID and TAC component tasks?

- Multimedia
- Streaming input
  - Can’t go back to reanalyze raw docs in previous data chunks
    - TA1 has access to TA2 KB encoding previously added KE’s
- Multiple hypotheses and interpretations
  - Expanded ontology to cover informational conflicts in scenario
  - TA1 outputs all possible extractions and interpretations, not just the most confident ones
  - TA1 extraction from data items may be conditioned on hypothesis

How is Task 2 different from Cold Start KBP?

- Multimedia
- Streaming input
  - TA2 has no access to raw data items to assist in fusing incoming KEs with existing KB; can only use what's represented in the incoming KE and existing KB
- Multiple hypotheses and interpretations
  - Expanded ontology to cover information conflicts in scenario
  - TA2 KB must maintain all possible KEs (even low-confidence KEs) in order to support creation of multiple hypotheses and disparate interpretations
  - TA2 KEs and confidences theoretically could be conditioned on hypothesis in future, but for 2018 the TA2 KB is independent of any "what if" hypotheses.

Evaluation by Assessment

- Evaluate using post-submission assessment and clustering of pooled mentions
  - To support evaluation of TA1 extraction conditioned on context, ground-truth must be conditioned on a small set of hypotheses, predetermined by LDC.
- Only targeted KEs (relevant to hypotheses) will be evaluated
- Only k highest-confidence mentions/justifications for each KE will be pooled and assessed
- LDC ***might*** provide exhaustive annotation of mentions of entities for a ***small*** set of documents, for gold-standard based “NER” evaluation

AIDA Evaluation Schedule

- 3 18-month phases
- January 2018 kick-off

- ~Sept 2018: Eval Pilot
- ~May 2019: Eval 1 (Phase 1)
- ~Nov 2020: Eval 2 (Phase 2)
- ~May 2022: Eval 3 (Phase 3)

TAC 2018 Streaming MM KBP Pilot Evaluation Schedule

- Sample/training/eval data release:
  - ~January: scenario and 3 mostly labeled topics for training; all 100K unlabeled docs for the scenario (foreign languages announced at this time)
  - ~April: 3 additional labeled topics for training
  - ~September: 6 “evaluation” topics
- Early September (?): Task 1 evaluation window
- Mid September (?): Task 2 evaluation window