Overview of the 2010 Text Analysis Conference

Sponsored by:

Hoa Trang Dang
National Institute of Standards and Technology

TAC Goals

- To promote research in NLP based on large common test collections
- To improve evaluation methodologies and measures for NLP
- To build test collections that evolve to anticipate the evaluation needs of modern NLP systems
- To increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas
- To speed transfer of technology from research labs into commercial products

Features of TAC

- Component evaluations situated within context of end-user tasks (e.g., summarization, knowledge base population)
  - opportunity to test components in end-user tasks
- Test common techniques across tracks
- Small number of tracks
  - critical mass of participants per track
  - sufficient resources per track (data, assessing, technical support)
- Leverage shared resources across tracks (organizational infrastructure, data, assessing, tools)

Track Participants

- Track Organizers
  - KBP: *Ralph Grishman, Heng Ji*, Paul McNamee, Boyan Onyshkevych; LDC data providers
  - RTE: *Luisa Bentivogli*, *Danilo Giampiccolo*, Peter Clark, Ido Dagan; with support from Pascal-2 Network of Excellence
  - Summarization: *Karolina Owczarzak*
- Annotators/assessors from LDC, CELCT, NIST
- 61 Teams
  - 18 countries
  - 5 continents (23 Asia, 17 N. America, 16 Europe, ...)

Overview

- Knowledge Base Population Track (KBP)
  - Entity-Linking Tasks (with/without wikipedia text)
  - Slot-Filling Tasks (known/surprise slots)
- Summarization Track
  - Guided (Update) Summarization Task
  - Automatically Evaluating Summaries of Peers (AESOP)
- Recognizing Textual Entailment Track (RTE-6)
  - Main and Novelty-Detection Tasks (Summarization setting)
  - KBP Validation Pilot (KBP slot-filling setting)

Overview

- Knowledge Base Population Track (KBP)
  - Entity-Linking Tasks (with/without wikipedia text)
  - Slot-Filling Tasks (known/surprise slots)
- Summarization Track
  - Guided (Update) Summarization Task
  - Automatically Evaluating Summaries of Peers (AESOP)
- Recognizing Textual Entailment Track (RTE-6)
  - Main and Novelty-Detection Tasks (Summarization setting)
  - KBP Validation Pilot (KBP slot-filling setting)

Knowledge Base Population Track

- Goal: Augment a reference knowledge base (KB) with info about target entities as found in a diverse collection of documents
- Reference KB: Oct 2008 Wikipedia snapshot. Each KB node corresponds to a Wikipedia and contains:
  - Infobox
  - Wiki_text (free text not in infobox)
- Source document collection: 1.8 million documents
  - 1.3 million newswire
  - 500 Web and other docs
- Two basic tasks:
  - Entity-linking: grounding entity mentions in docs to KB nodes
  - Slot-filling: Learning attributes about target entities

Entity-Linking Task

- Task: link each query (name + document) to a node in the KB, or NIL if not in KB
- Evaluation Metric: Accuracy (averaged over all queries)
- Evaluation Results:
  - Entity-Linking
    - Participants: 16 teams
    - Highest System Accuracy: 86%
    - Human Accuracy (sample): ~90%
  - Optional Entity-Linking (no wikitext)
    - Participants: 7 teams
    - Highest System Accuracy: 78%

Slot-Filling Task

- Task: given target entity and predefined slots for each entity type (PER, ORG), return all slot fillers for that entity, and a supporting document for each filler
- Response format and evaluation based on TREC-QA pooling methodology for evaluating list questions
- Evaluation:
  - Set of [docid, answer-string] pairs for each target entity and attribute (slot)
  - Each pair judged as one of {wrong, inexact, redundant, correct}
  - Correct pairs grouped into equivalence classes (entities)
  - Recall: number of correct equivalence classes returned / number of known equivalence classes
  - Precision: number of correct equivalence classes returned / number of [docid, answer-string] pairs returned
  - F1 = (P*R)/(R+P)

Slot-Filling Evaluation Results

- Regular Slot-Filling

Participants:	15 teams
Highest System F1:	65.78
2nd Highest System F1:	29.15
Human F1:	61.06

- Surprise Slot-Filling (4 new slots, <= 4 days to customize system)

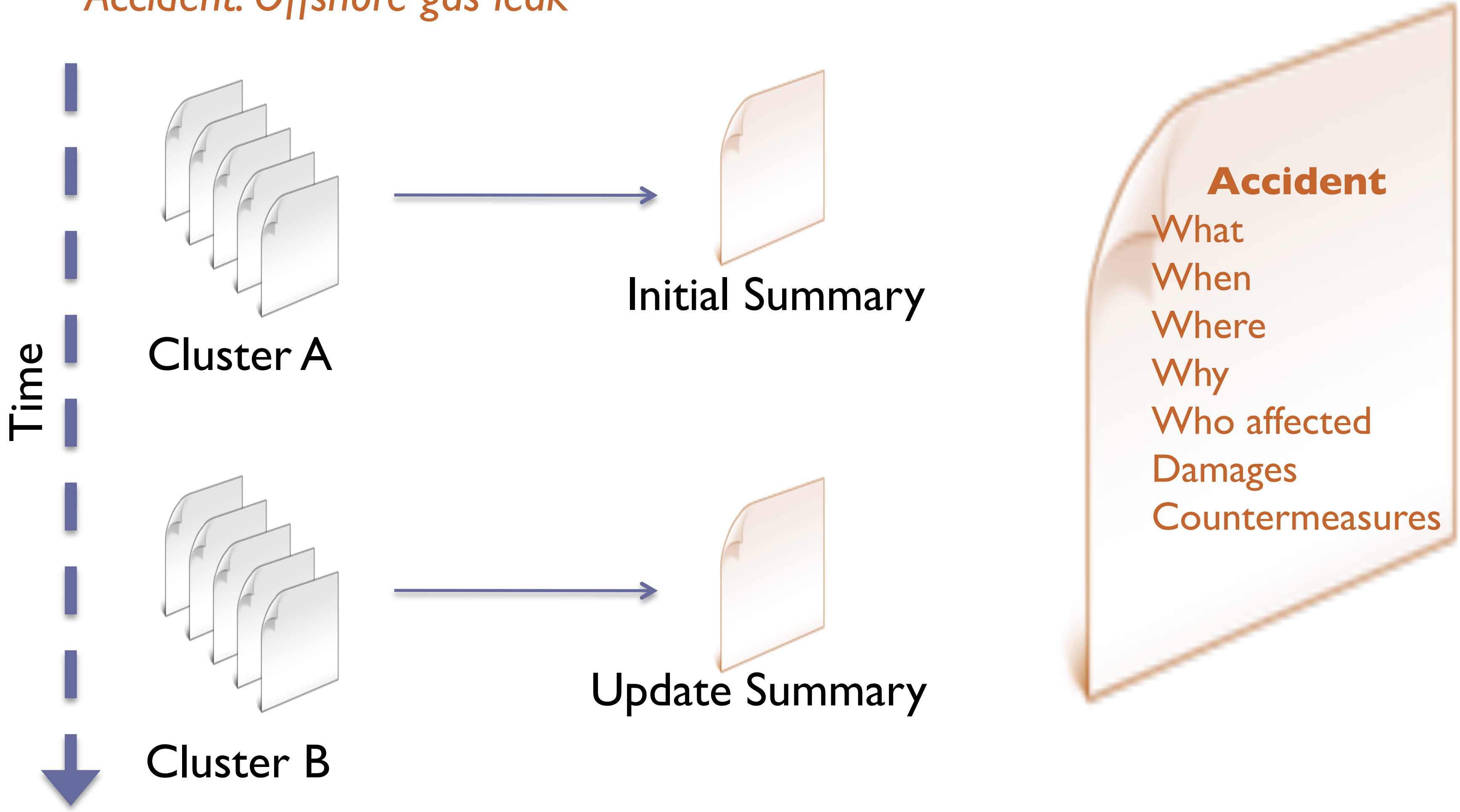
Participants:	5 teams
Highest System F1:	69.56 (semi-automatic, 99 hrs)
2nd Highest System F1:	33.06 (34 hrs)
Human F1:	56.80

Overview

- Knowledge Base Population Track (KBP)
  - Entity-Linking Tasks (with/without wikipedia text)
  - Slot-Filling Tasks (known/surprise slots)
- Summarization Track
  - Guided (Update) Summarization Task
  - Automatically Evaluating Summaries of Peers (AESOP)
- Recognizing Textual Entailment Track (RTE-6)
  - Main and Novelty-Detection Tasks (Summarization setting)
  - KBP Validation Pilot (KBP slot-filling setting)

Guided Update Summarization Task

Accident: Offshore gas leak



Time

Cluster A

Initial Summary

Cluster B

Update Summary

Accident

What

When

Where

Why

Who affected

Damages

Countermeasures

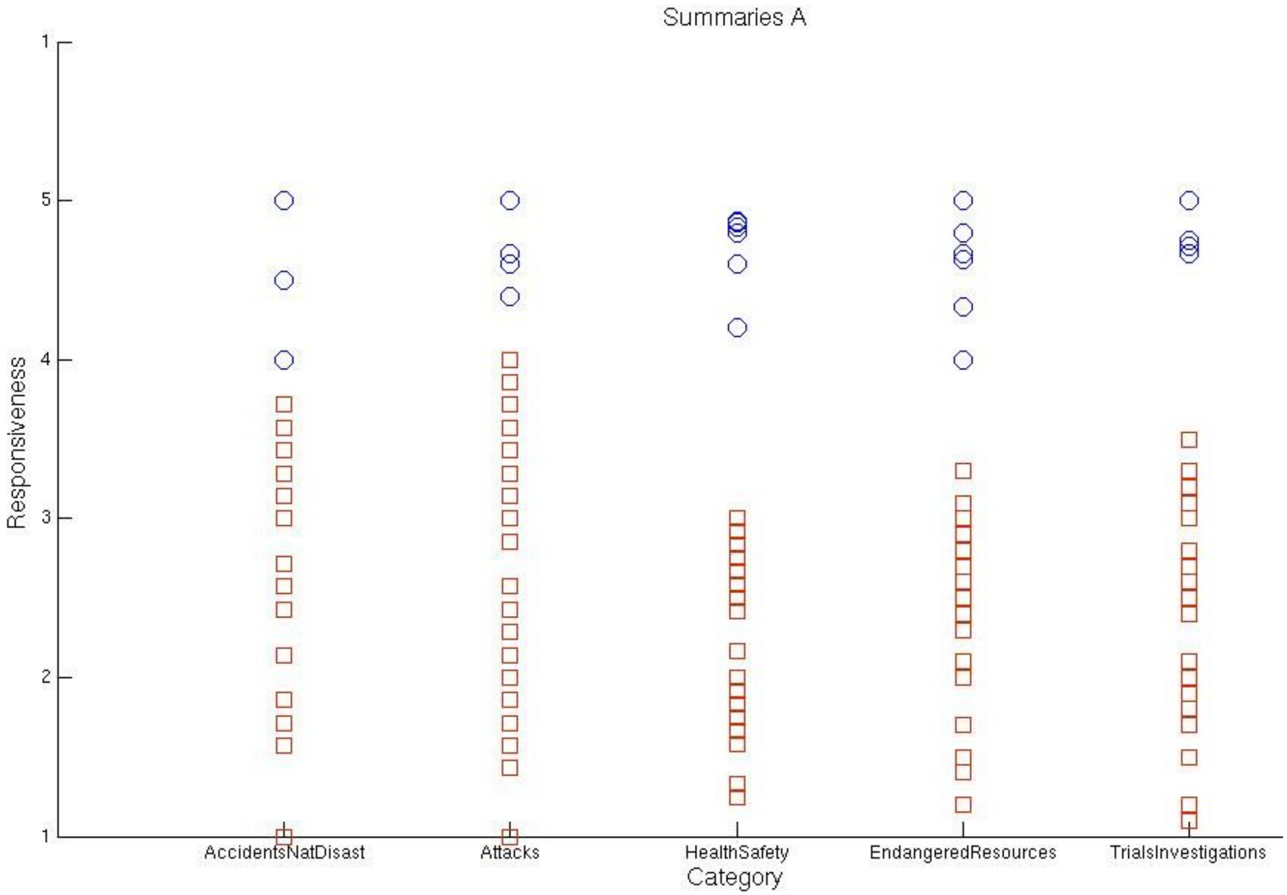
Summarization Topic Categories and Aspects

Accidents, Natural Disasters	Attacks	Health and Safety	Endangered Resources	Investigations and Trials
What	What	What	What	Who
When	When	Who affected	Importance	Investigators
Where	Where	How	Threats	Why
Why	Perpetrators	Why	Countermeasures	Charges
Who affected	Why	Countermeasures		Plead
Damages	Who affected			Sentence
Countermeasures	Damages			
	Countermeasures			
Other	Other	Other	Other	Other

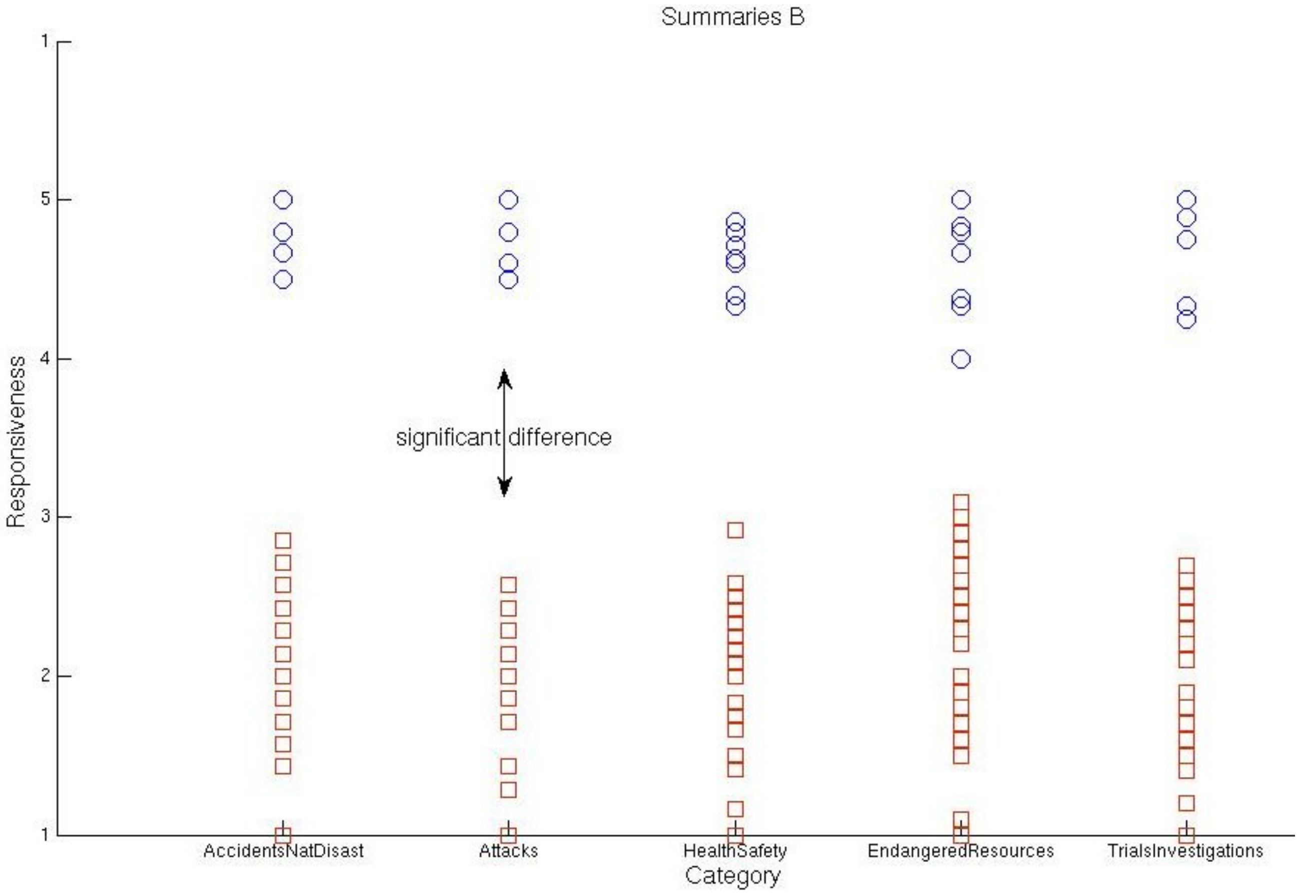
Guided Summarization Task

- Human Abstractors: 8 NIST assessors writing model (reference) summaries
- Participants: 23 teams; 41 runs (summarizers), plus 2 baselines
- Evaluation:
  - Pyramid Evaluation of summary content (Passonneau et al., 2005), overlap with human-authored summaries
    - multiple human summaries
    - summary content unit ("nugget") weighted by number of human summaries it appears in
  - Overall Readability
  - Overall Responsiveness (Readability and responsiveness to required aspects for the topic)

Responsiveness by Category (Initial Summaries)



Responsiveness by Category (Update Summaries)



Overview

- Knowledge Base Population Track (KBP)
  - Entity-Linking Tasks (with/without wikipedia text)
  - Slot-Filling Tasks (known/surprise slots)
- Summarization Track
  - Guided (Update) Summarization Task
  - Automatically Evaluating Summaries of Peers (AESOP)
- Recognizing Textual Entailment Track (RTE-6)
  - Main and Novelty-Detection Tasks (Summarization setting)
  - KBP Validation Pilot (KBP slot-filling setting)

Automatically Evaluating Summaries of Peers

- Goal: Develop automatic metrics that emulate manual metrics measuring quality of summary content (Responsiveness, Pyramid)
- Participants: 9 teams, 24 AESOP metrics
- Evaluation:
  - Summarizer-level correlations with manual metrics
    - High summarizer-level correlation between AESOP metrics and manual metrics
  - Discriminative power between summarizers as compared to discriminative power of manual metrics
    - High similarity in discriminative power of manual metrics and some participants' AESOP metrics

Overview

- Knowledge Base Population Track (KBP)
  - Entity-Linking Tasks (with/without wikipedia text)
  - Slot-Filling Tasks (known/surprise slots)
- Summarization Track
  - Guided (Update) Summarization Task
  - Automatically Evaluating Summaries of Peers (AESOP)
- Recognizing Textual Entailment Track (RTE-6)
  - Main and Novelty-Detection Tasks (Summarization setting)
  - KBP Validation Pilot (KBP slot-filling setting)

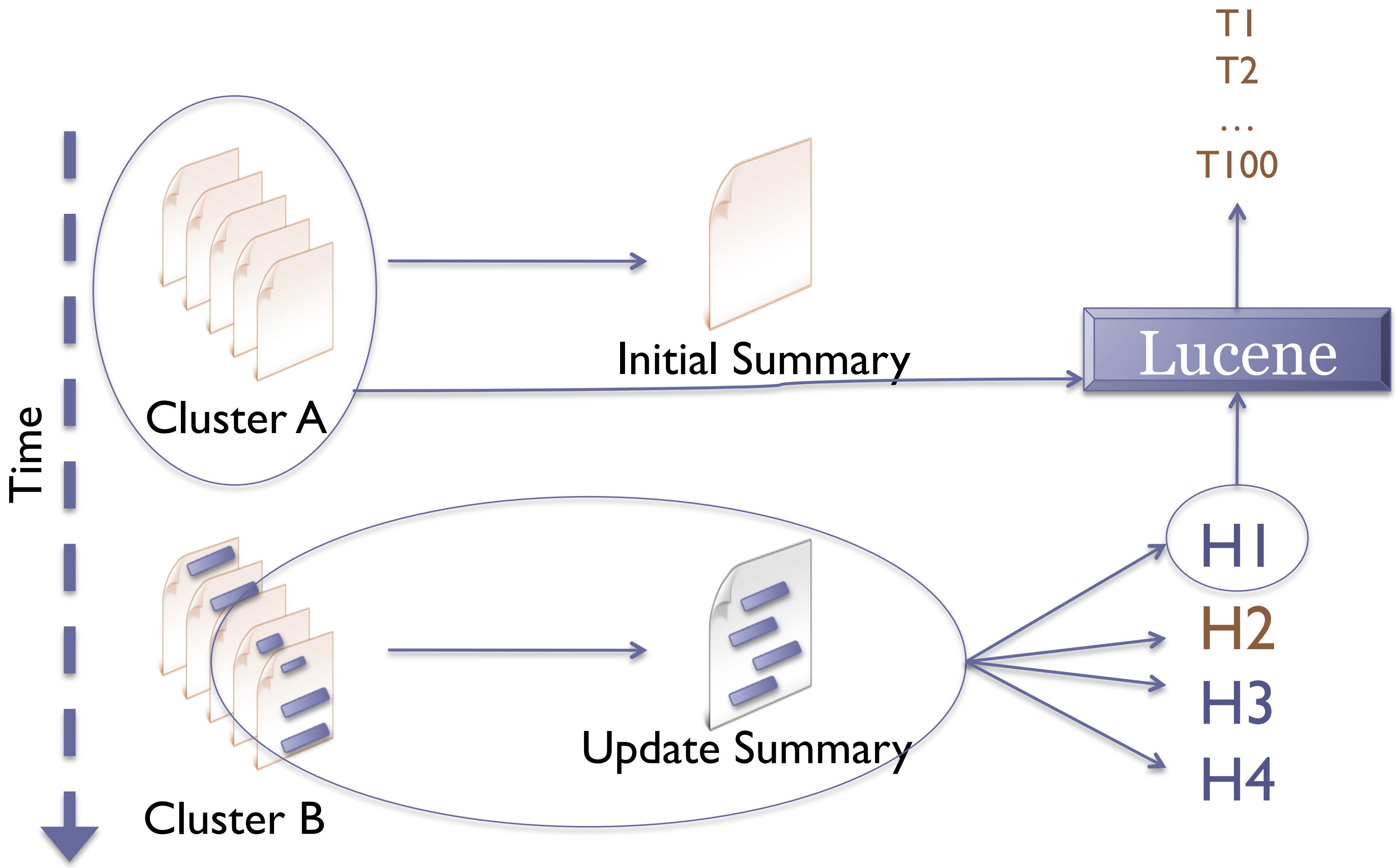
Recognizing Textual Entailment Track (RTE-6)

- Textual entailment is a directional relation between two text fragments: **T**(ext) and **H**(ypothesis)
  - *T Entails H if a human reading T would infer that H is most likely true*
    - *T: The knifeman who carried out Japan's worst killing rampage in central Tokyo yesterday, killing 7 people, may have been planning the attack for months.*
    - *H: Seven people were killed by a knifeman in Tokyo.* **YES**
- RTE-6 tasks situated in and supporting TAC applications
  - Summarization Setting - Main Task, Novelty detection
  - KBP Setting – Validation of KBP slot fillers
- Challenge: judging entailment in larger context of one or more documents, interpreting explicit and implicit references to entities, places, dates, events

Overview

- Knowledge Base Population Track (KBP)
  - Entity-Linking Tasks (with/without wikipedia text)
  - Slot-Filling Tasks (known/surprise slots)
- Summarization Track
  - Guided (Update) Summarization Task
  - Automatically Evaluating Summaries of Peers (AESOP)
- Recognizing Textual Entailment Track (RTE-6)
  - Main and Novelty-Detection Tasks (Summarization setting)
  - KBP Validation Pilot (KBP slot-filling setting)

RTE: Update Summarization Setting



T-H Pairs from Summaries and Documents

- Extractive update summary/docid AFP_ENG_20050428.0315: “Suspected Muslim rebels killed three policemen, a state political party member and two others in the first attacks on police in Kashmir since Indian and Pakistan leaders met two week's ago.”
  - H610: Suspected Muslim rebels killed three policemen
  - H611: Suspected Muslim rebels killed a state political party member
  - H605: Indian and Pakistan leaders met in April 2005.
- For each H, up to 100 candidate sentences retrieved by Lucene from Cluster A, using H as query
- Task: For each H, retrieve all candidate sentences T such that T entails H (T and H interpreted in context of entire cluster of *documents)*

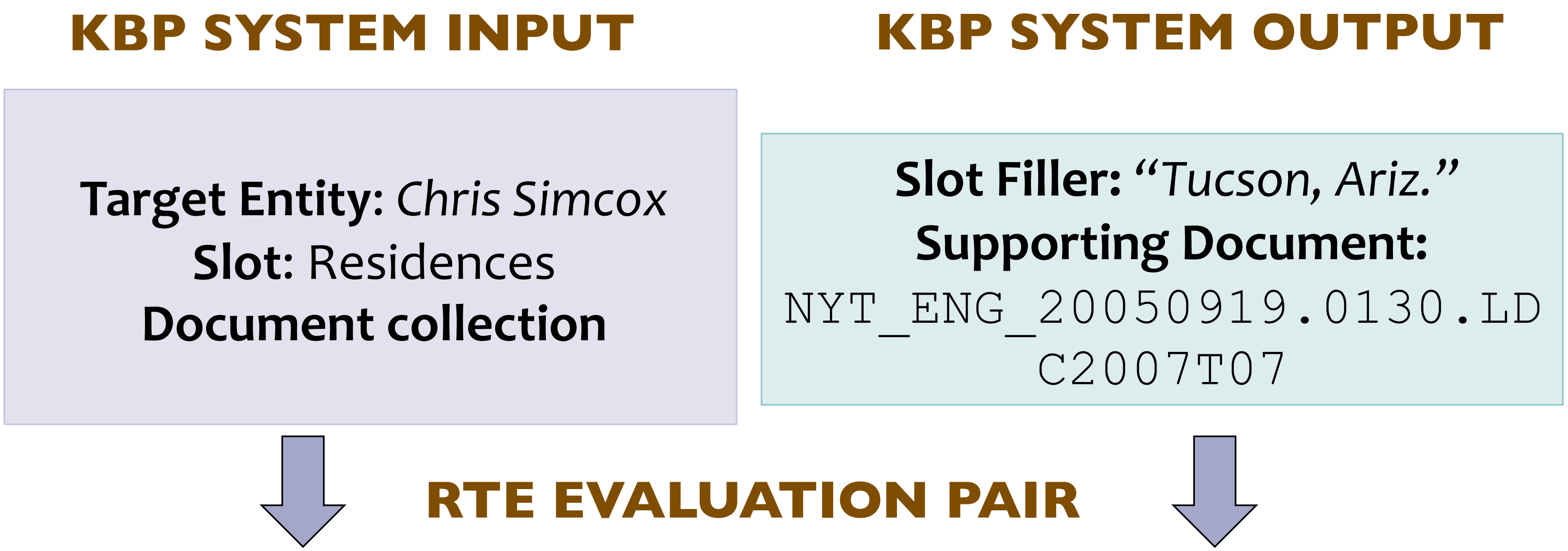
RTE in Summarization Setting

- Main Task
  - Evaluation Metrics: micro-averaged P/R/F1 on correctly retrieved entailing sentences
  - Participants: 18 teams
  - Evaluation Results:
    - Best Run F1: 48.01
    - Lucene5 Baseline F1: 34.63
- Novelty Detection subtask: no sentences entailing H ⇔ novel H
  - Evaluation: P/R/F1 on novel H's detected
  - Participants: 9 teams
  - Evaluation Results:
    - Best Run F1: 82.91
    - Baseline (all novel) F1: 66.89

Overview

- Knowledge Base Population Track (KBP)
  - Entity-Linking Tasks (with/without wikipedia text)
  - Slot-Filling Tasks (known/surprise slots)
- Summarization Track
  - Guided (Update) Summarization Task
  - Automatically Evaluating Summaries of Peers (AESOP)
- Recognizing Textual Entailment Track (RTE-6)
  - Main and Novelty-Detection Tasks (Summarization setting)
  - KBP Validation Pilot (KBP slot-filling setting)

RTE KBP Validation: Creating T-H Pairs



KBP SYSTEM INPUT

Target Entity: *Chris Simcox*
Slot: Residences
Document collection

KBP SYSTEM OUTPUT

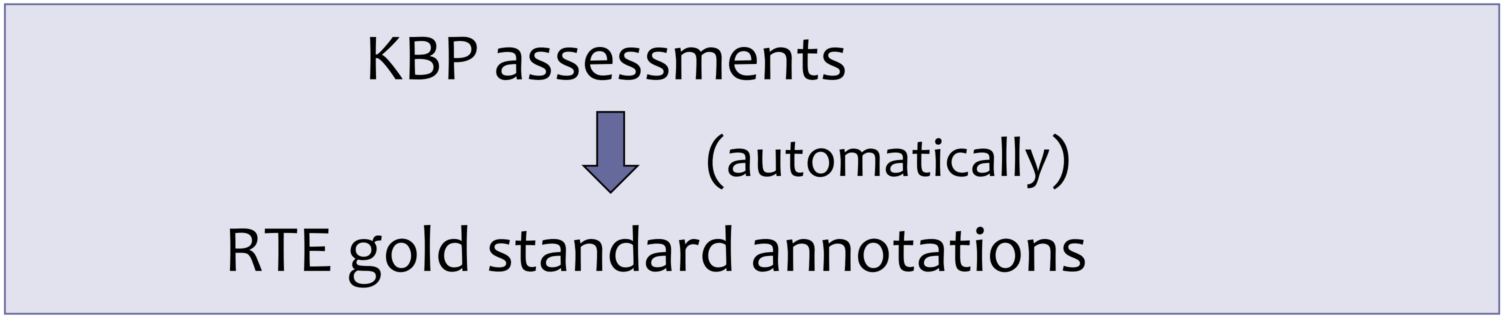
Slot Filler: *“Tucson, Ariz.”*
Supporting Document:
`NYT_ENG_20050919.0130.LDC2007T07`

RTE EVALUATION PAIR

T: `NYT_ENG_20050919.0130.LDC2007T07`

H:	H1:	*Chris Simcox lives in Tucson, Ariz.*
	H2:	*Chris Simcox has residence in Tucson, Ariz.*
	H3:	*Tucson, Ariz. is the place of residence of Chris Simcox*
	H4:	*Chris Simcox resides in Tucson, Ariz.*
	H5:	*Chris Simcox’s home is in Tucson, Ariz.*

RTE KBP Validation: Creating the Gold Standard



KBP assessments

(automatically)

RTE gold standard annotations

KBP JUDGMENTS (4-valued)	ENTAILMENT VALUES (2-valued)
Correct	YES
Redundant	YES
Wrong	NO
Inexact	(not included)

RTE KBP Validation Pilot

- Evaluation Metrics: micro-averaged P/R/F1 on T-H pairs
- Baseline: All T's classified as entailing the corresponding H
  - Reflects cumulative performance of all KBP slot-filling systems
  - Precision is the percentage of entailing pairs in test set
- Participants: 3 teams
- Evaluation Results:
  - Best Run F1: 25.5 (33.07 if tailored to slots)
  - Baseline F1: 16.13

TAC 2011 Tracks

1. RTE
2. KBP (+ multilingual)
3. Summarization

- Come to the track planning sessions!