IBM T.J. Watson Research Center

IBM Marvel for TRECVID06 Automatic Search

Intelligent Information Management Dept.
IBM Thomas J. Watson Research Center

Contact: Paul Natsev <natsev@us.ibm.com>

deeper

This material is based upon work funded in part by the U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Government.

November 14th, 2006
Gaithersburg, MD

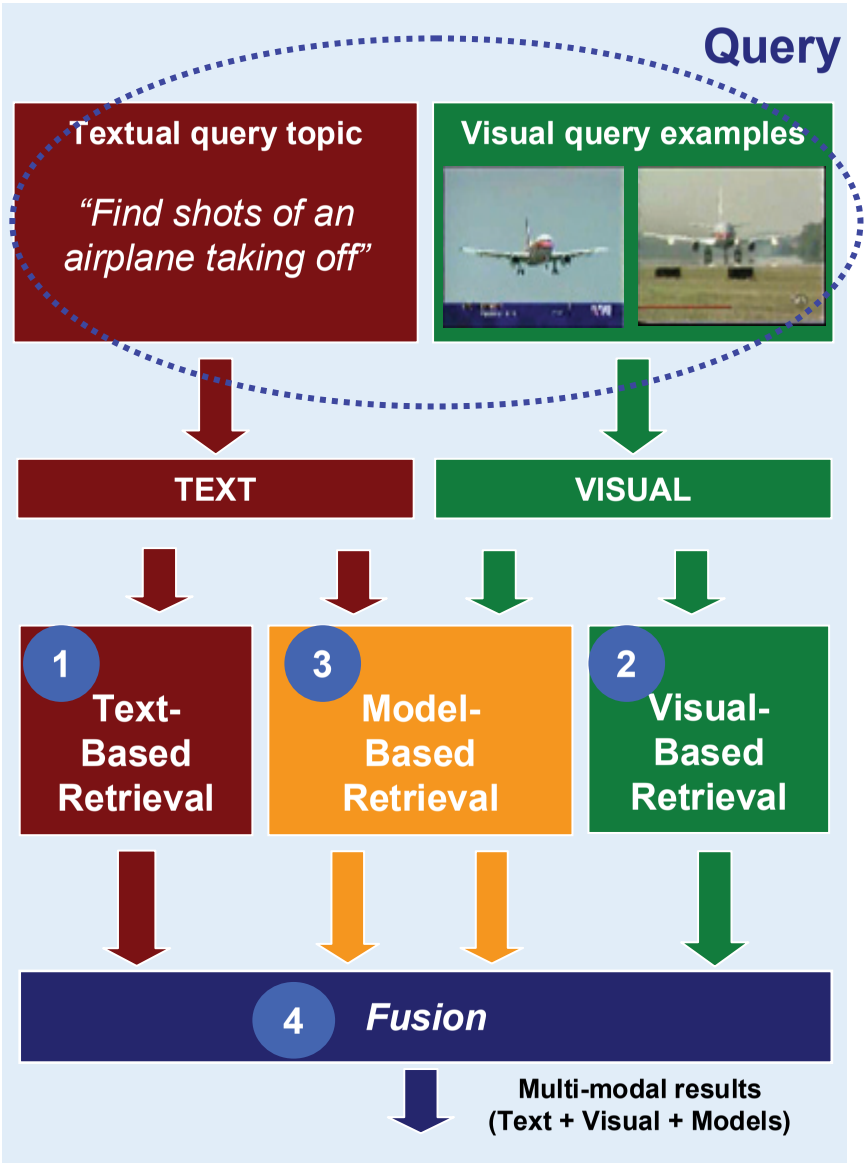
Acknowledgments

- IBM Research Intelligent Information Management Team
  - Apostol (Paul) Natsev
  - Jelena Tesic
  - John R. Smith
  - Lexing Xie
  - Milind Naphade
  - Murray Campbell
  - Quoc-Bao Nguen
  - Shahram Ebadollahi
  - Alex Haubold (Columbia U.)
  - Dhiraj Joshi (Penn. State)
  - Joachim Seidl (U. Klagenfurt, Austria)
- IBM Research Knowledge Structures Group (PIQUANT-II Q&A system)
  - Jennifer Chu-Carroll, John Prager, Pablo Duboue
- Columbia University (Story boundaries)
  - Lyndon Kennedy, Winston Hsu, Shih-Fu Chang
- National University of Singapore (Phrase-aligned and speaker-aligned transcripts)
  - Shi-Yong Neo, Tat-Seng Chua

Outline

- System Overview
- Text-Based Retrieval
- Visual Content-Based Retrieval
- Concept Model-Based Retrieval
- Multi-modal Fusion
- Results and Conclusions

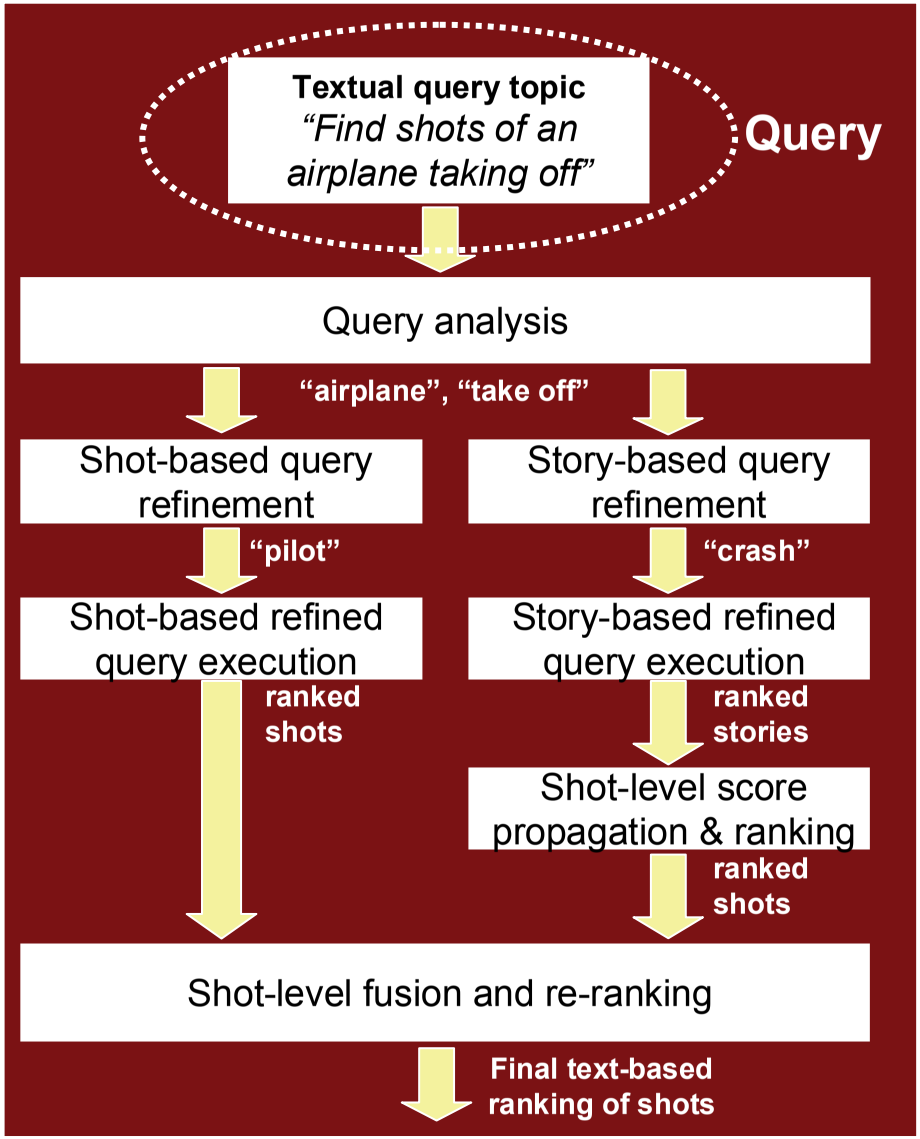
IBM Research Automatic Search System Overview



Approaches:

1. Text-based: story-based retrieval with automatic query refinement/re-ranking
2. Visual-based: light-weight learning (discriminative and nearest neighbor modeling) with smart sampling
3. Model-based: automatic query-to-model mapping based on query topic text or visual examples
4. Fusion:
   - Query independent
   - Query-class-dependent with soft or hard class membership

1 IBM Research Text Retrieval System



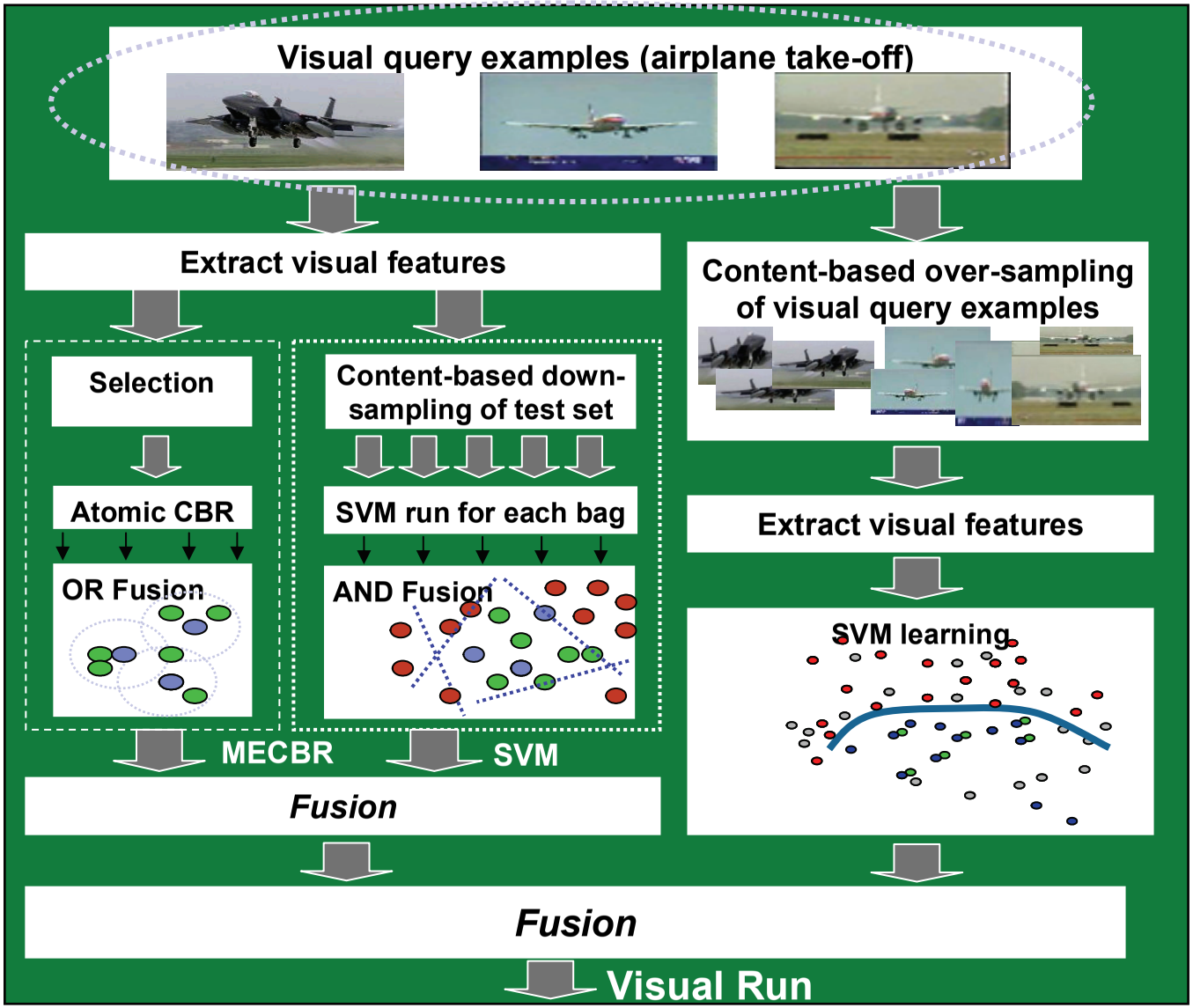
- Corpus Indexing
  - Shot-level ASR/MT documents
  - Story-level ASR/MT documents
  - Both aligned at phrase level
- Query analysis:
  - Tokenization, phrasing, stemming
  - Part-of-speech tagging & filtering
- Query refinement:
  - Pseudo-Relevance Feedback
- Query execution and fusion
  - IBM Semantic Search Engine (Juru)
  - Using TF*IDF-based retrieval
  - Fusion of shot- and story-based results

Performance (MAP):

- Shots: 0.041
- Stories: 0.032
- Fused: 0.052 (our best uni-modal run)

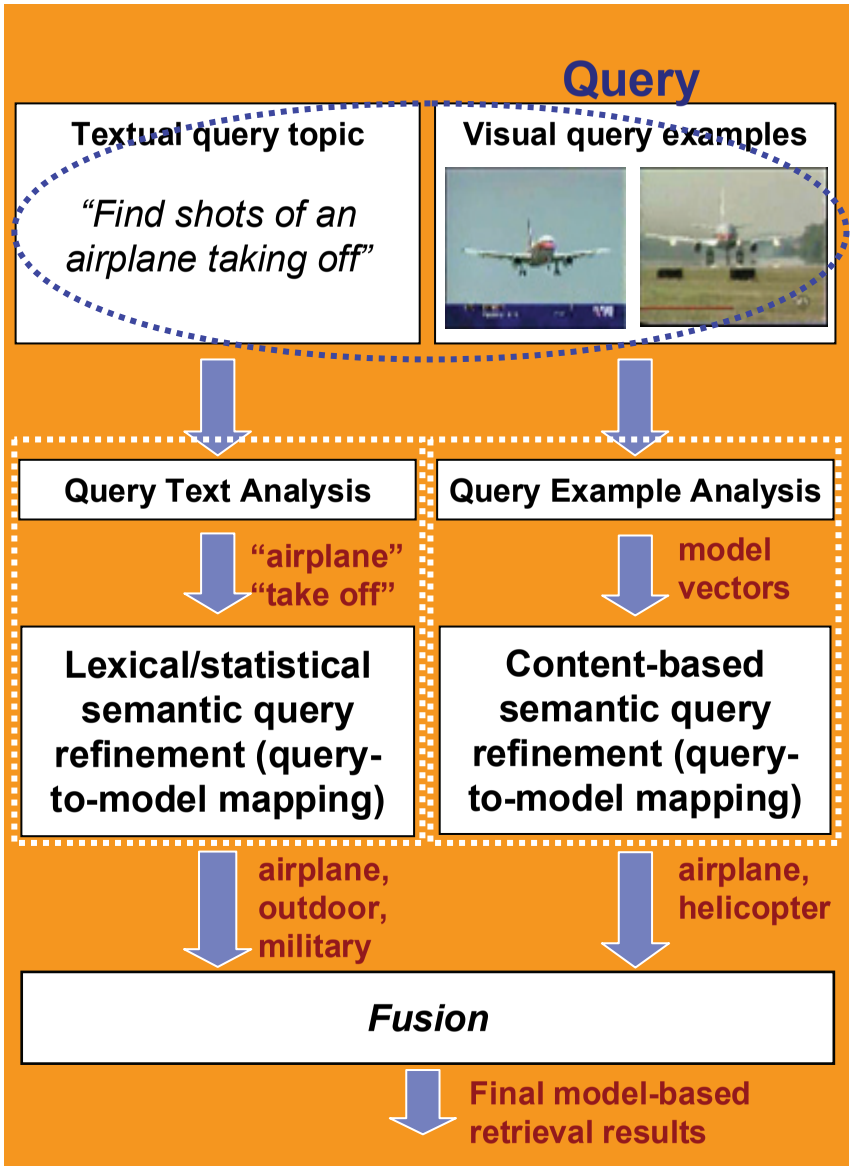
- See Volkmer & Natsev (ICME 2006)

3 IBM Research Visual Retrieval System



- **Unbalanced learning from multimedia data**
- Content-based down-sampling of test set to create negative examples
  - Fusion of two light-weight learning techniques: k-NN and SVM
- Content-based over-sampling of visual query to create positive examples
  - SVM learning
  - MAP: 0.0212
- See Natsev et al. (ACM MM 2005)

2 IBM Research Model-Based Retrieval System



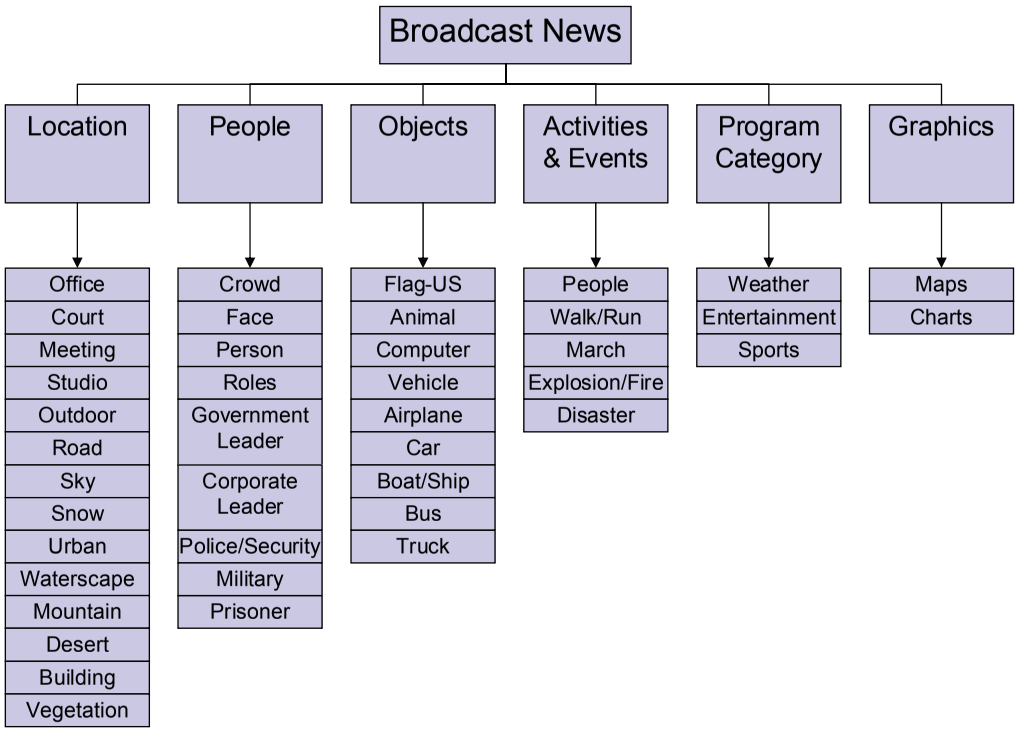
Approaches:

1. Query text-based: automatically map query topic text to concept models
   - Lexical (WordNet similarity)
   - Lexical (lexicon mapping)
   - Lexical (direct word match)
   - Statistical (co-occurrence-based)
   - Statistical (pseudo-RF)
2. Query example-based: automatically map query examples to concept models
   - Content-based (model vectors)
   - Content-based (feature selection)
3. Fusion:
   - Query independent (MAP 0.045)
   - Query-class-dependent

2 Model-Based Retrieval—Semantic Lexicon

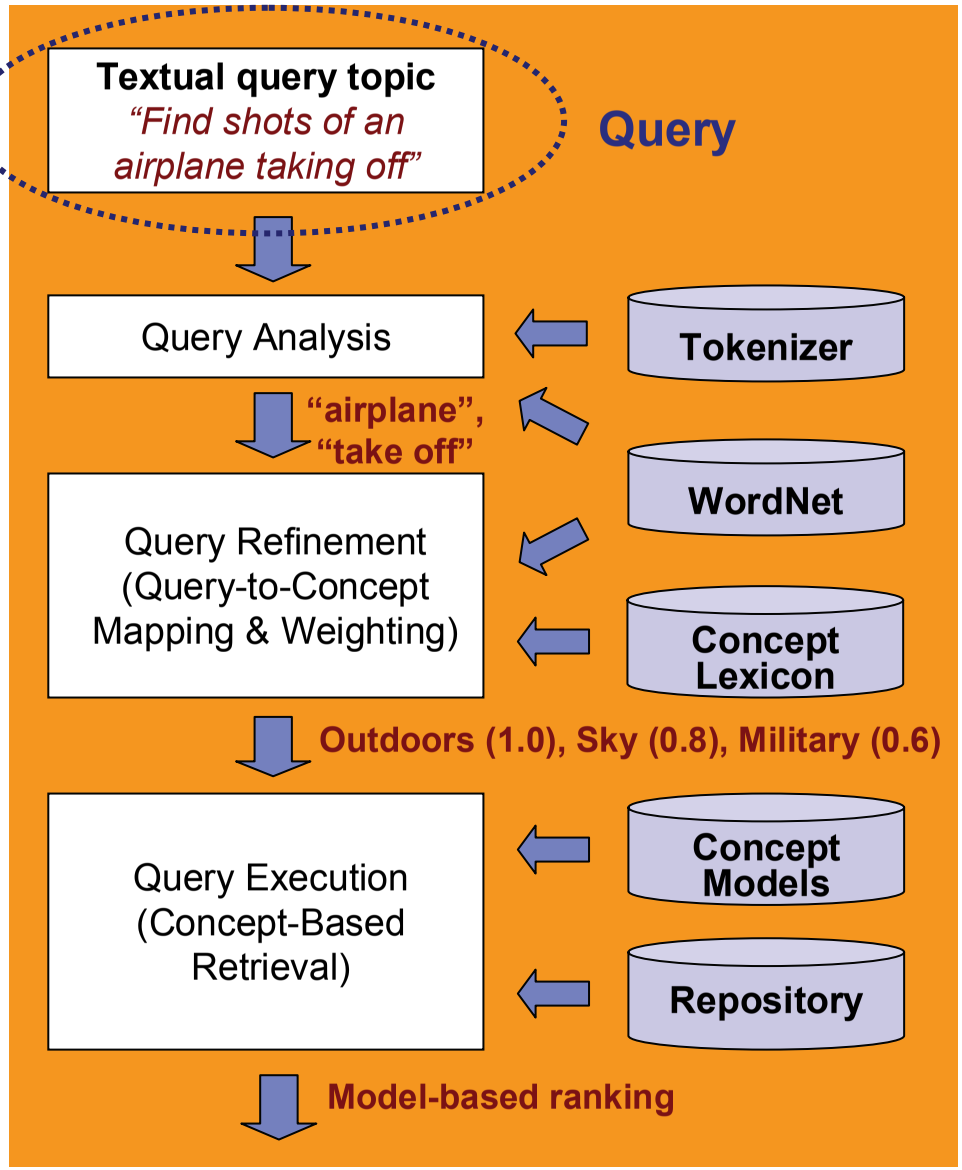
- Observations
  - Visual concepts can help refine query topics
  - Visual context can disambiguate word senses
- Idea
  - Leverage automatic visual concept detectors for semantic model-based query refinement

- Semantic Concept Lexicon
  - Hierarchy of 39 LSCOM-lite concepts
  - Statistical models based on visual features and machine learning
  - Each concept model described by several words/phrases from WordNet



Visual Concept	Representative words/phrases
Airplane	jet, aircraft, carrier, warplane
Natural disaster	disaster, earthquake, fire, flame, flood, hurricane, tornado, tsunami
Sports	sport, baseball, basketball, soccer, tennis, cricket, football, hockey, golf, game, match
US flag	American flag, stars and stripes

2 Model-Based Retrieval—Lexical (WordNet)



1. Query analysis
   - Tokens, stems, phrases
   - Stop word removal

2. Query refinement
   - Automatic mapping of query text to concept models & weights
   - Lexical approach based on WordNet Lesk similarity

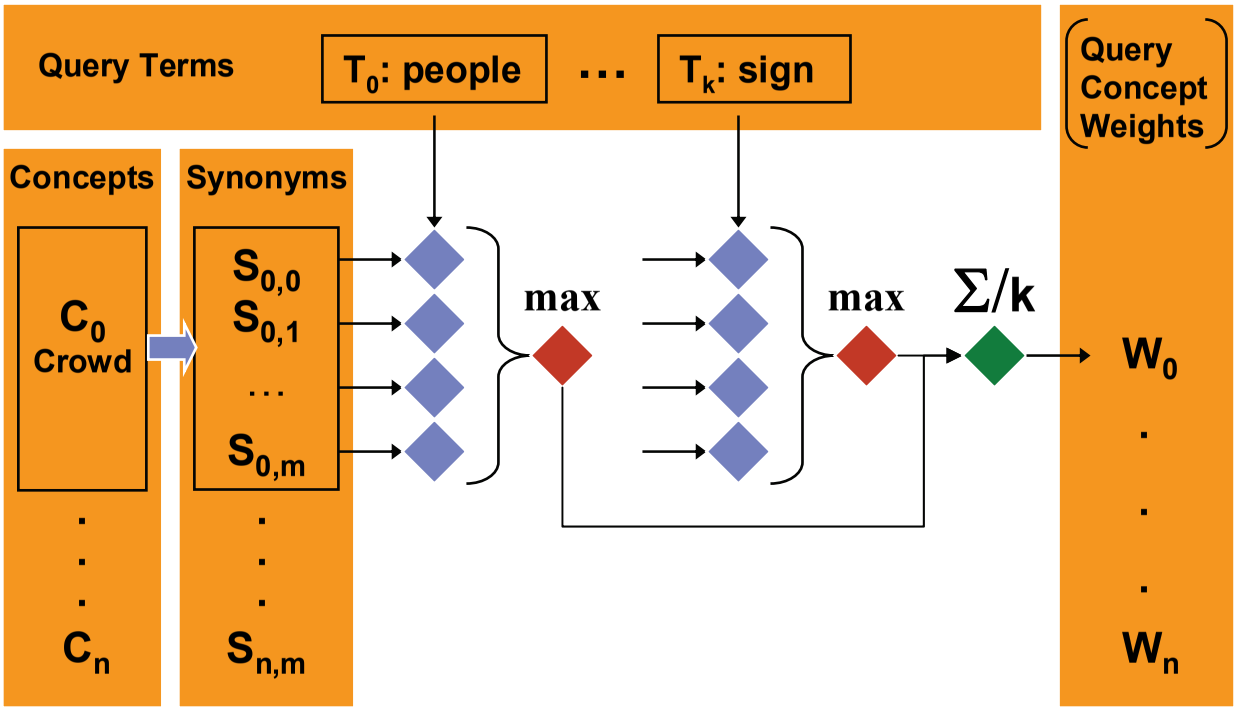
3. Query execution
   - Concept-based retrieval using statistical concept models
   - Weighted averaging model fusion

MAP: 0.029

See Haubold et al. (ICME 2006)

2 Lexical Query Refinement (Query-to-Concept Mapping)

- Semantic Relatedness
  - –WordNet Lesk similarity between Concept Synonyms and Query Terms
  - –max(Lesk) between Concept and Query Term determines closest relatedness
  - –Normalized sum over query terms establishes Query Concept Weight vector



Query: "people with banners or signs"	
T_0: *people*	*people-marching*: 4737, *crowd*: 4737, …
T_1: *banner*	*people-marching*: 216, *US flag*: 151, …
T_2: *sign*	*building*: 5361, *waterscape/front*: 531, …
Query-to-concept mapping & weights: *Building*: 1879.7 *People/marching*: 1685.7 *Crowd*: 1625.3 …	

2 Query Execution (Model-Based Retrieval of Shots)

- Fusion of weight vectors
  - Query concept weight vector
  - Shot concept confidence vector
- Final shot score
  - $\Sigma(\Pi)$ between query and shot concept vectors
  - Sorted shot scores: ranked list of results

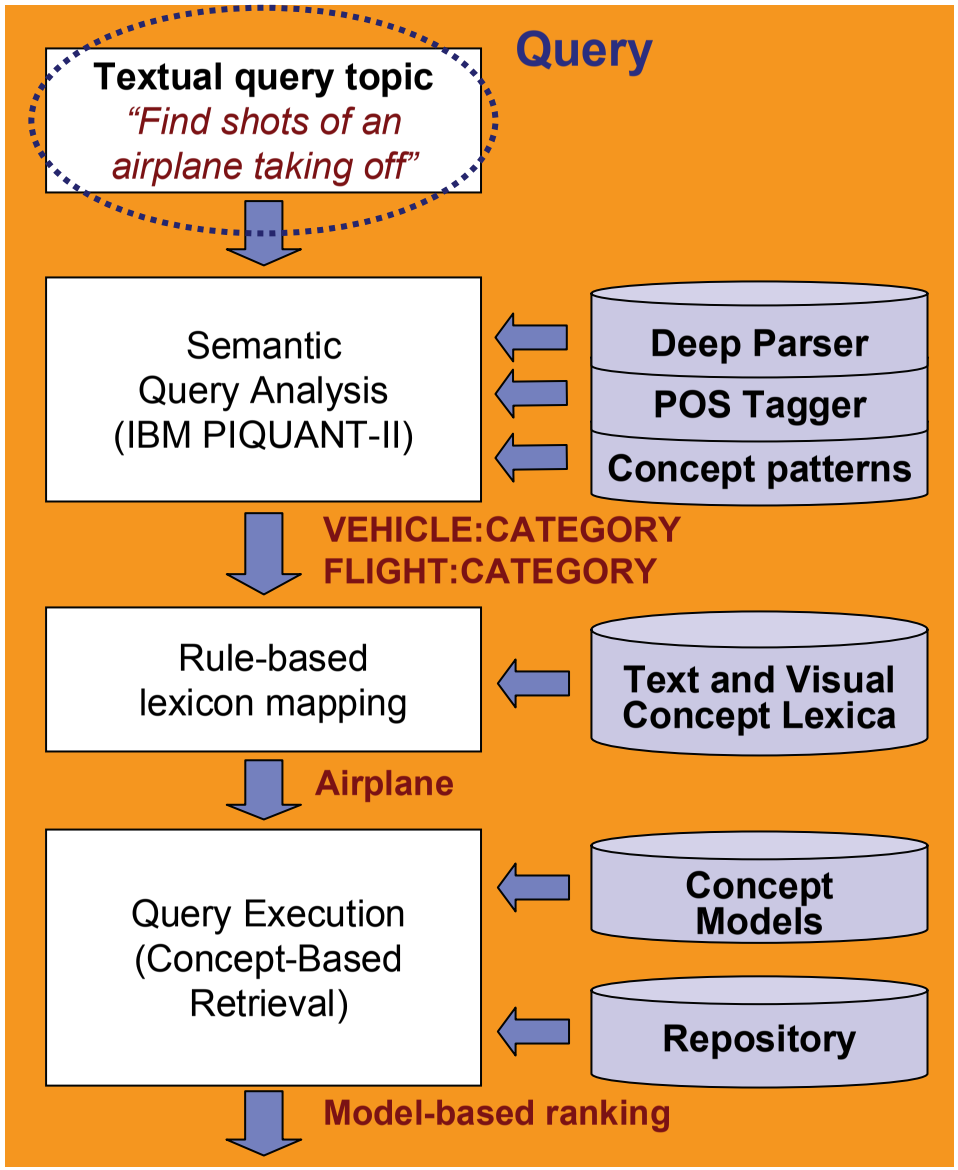
$$\text{Query} < W_0 , W_1 , W_2 , W_3 , \ldots , W_k >$$
$$\text{Shot}_0 < C_0 , C_1 , C_2 , C_3 , \ldots , C_k >$$

$\Sigma(\Pi)$: S_0 Score

$$\text{Query} < W_0 , W_1 , W_2 , W_3 , \ldots , W_k >$$
$$\text{Shot}_n < C_0 , C_1 , C_2 , C_3 , \ldots , C_k >$$

$\Sigma(\Pi)$: S_n Score

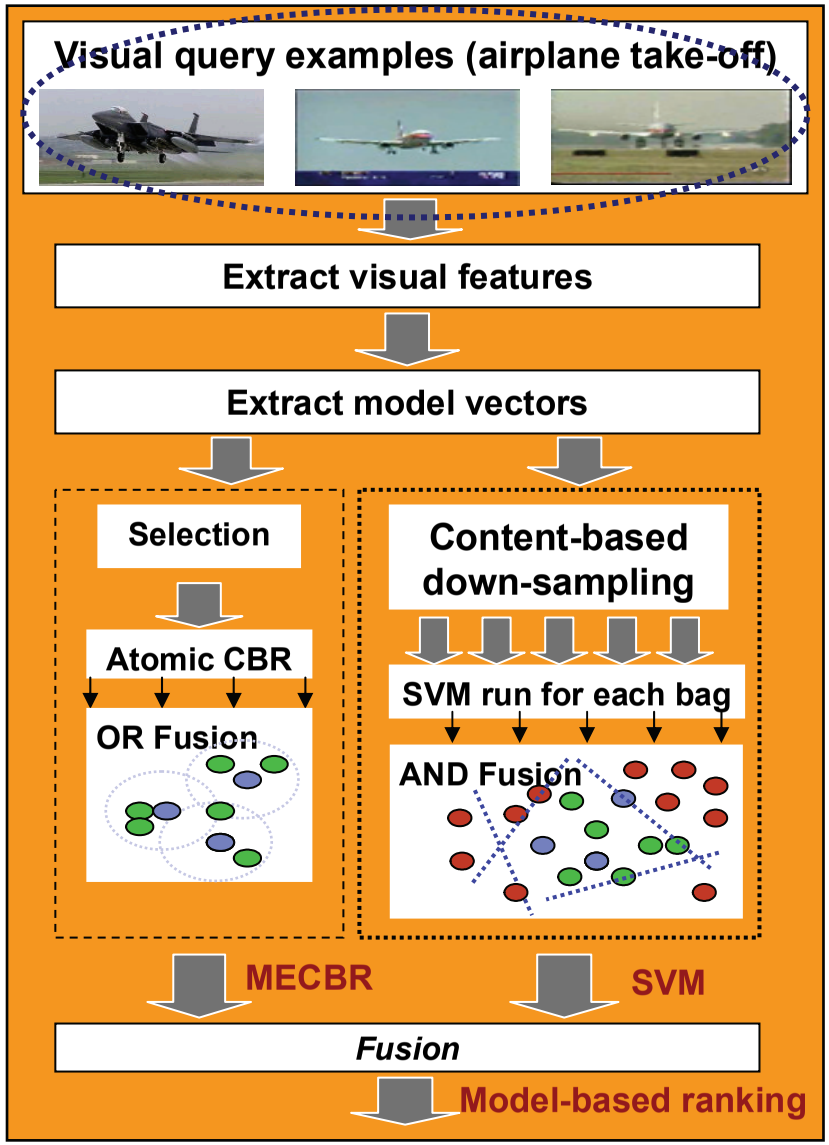
2 Model-Based Retrieval—Lexicon Mapping



1. Semantic query analysis
   - Semantic annotations of 100+ named and unnamed entities in text (people, places, events, etc.)
   - IBM PIQUANT-II Q&A engine
2. Rule-based lexicon mapping
   - Manually defined rules for mapping text annotation types to LSCOM-lite visual concepts
3. Query execution
   - Concept-based retrieval using statistical concept models
   - Weighted averaging model fusion

Performance:
- MAP: 0.018
- QRY 196 AP: 0.251

2 Model-Based Retrieval—Content-Based Approach

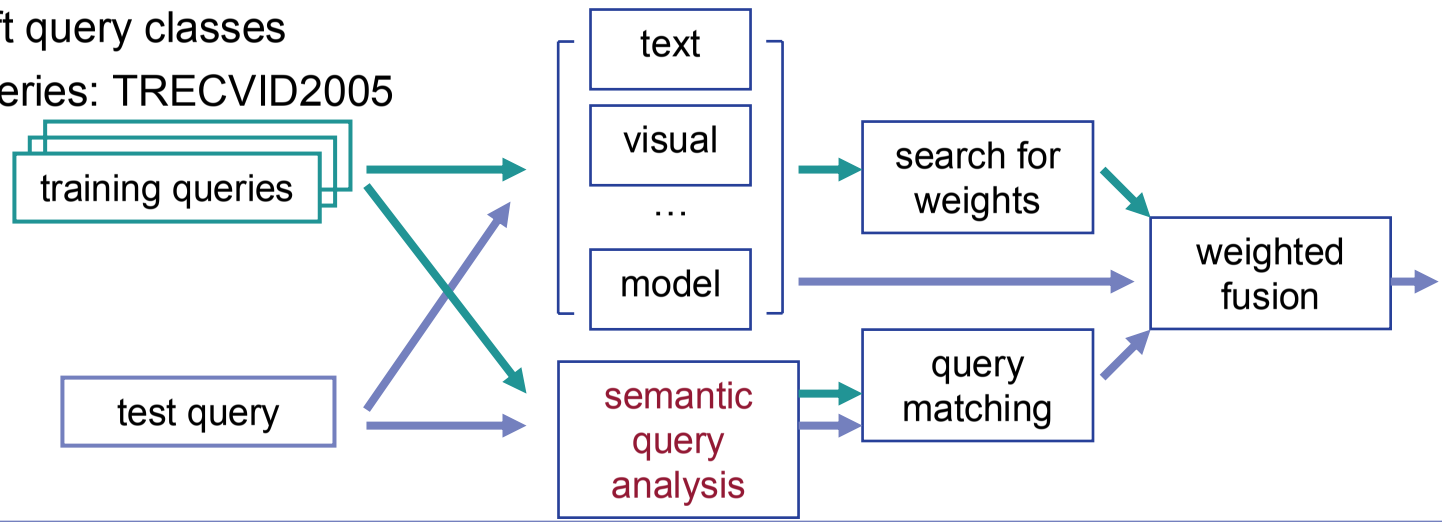


- Semantic model vector feature
  - 39-dimensional vector of confidences
  - Dimensions correspond to LSCOM-lite
- Learning technique
  - Fusion of two light-weight learning techniques: k-NN and SVM
- Sampling technique:
  - Sample pseudo-negative examples using coherent cluster centroids
- Best individual model-based approach
  - MAP: 0.034
  - QRY 195 AP: 0.5974

- See Natsev et al. (ACM MM 2005)

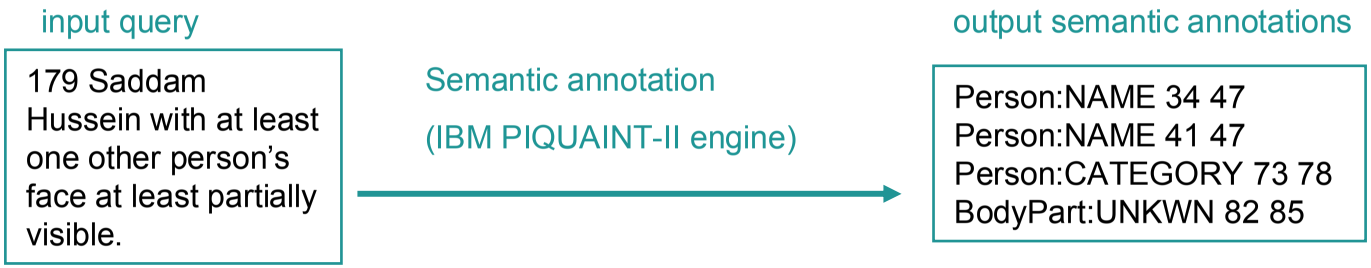
4 Multi-modal Fusion

- The problem
  - Text-, visual- and model-based retrieval are each good at finding certain things
    - Text: named people, other named entity
    - Visual: semantic scenes consistent in color or layout, e.g., sports, weather
    - Concept Models: non-specific queries, e.g. protest, boat, fire
  - Averaging the retrieval models helps [IBM TRECVID 2005]
  - Query-class or query-cluster dependent fusion helps more [CMU, NUS, Columbia]
- Our approach
  - Weighted linear combination for fusion
  - Semantic query analysis for query matching (based on PIQUANT-II Q&A technology)
  - Hard vs. soft query classes
  - Training Queries: TRECVID2005



Query Dependent Fusion Components

- Semantic query analysis



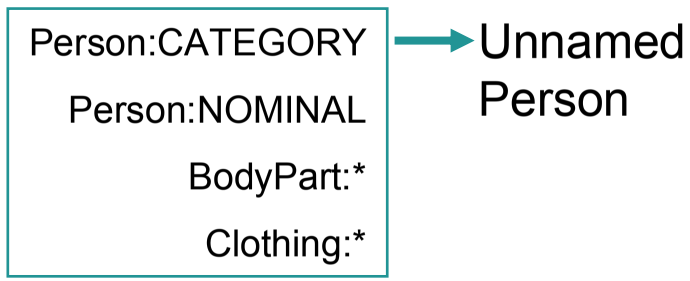
- IBM PIQUANT-II semantic annotation engine and type system
  - Designed for intelligence question-answering
  - Defined more than 100 semantic entities over text
    - (Named/generic) person, object, event, location, etc.
- Query class / query-component mapping

(a) 4 query classes:

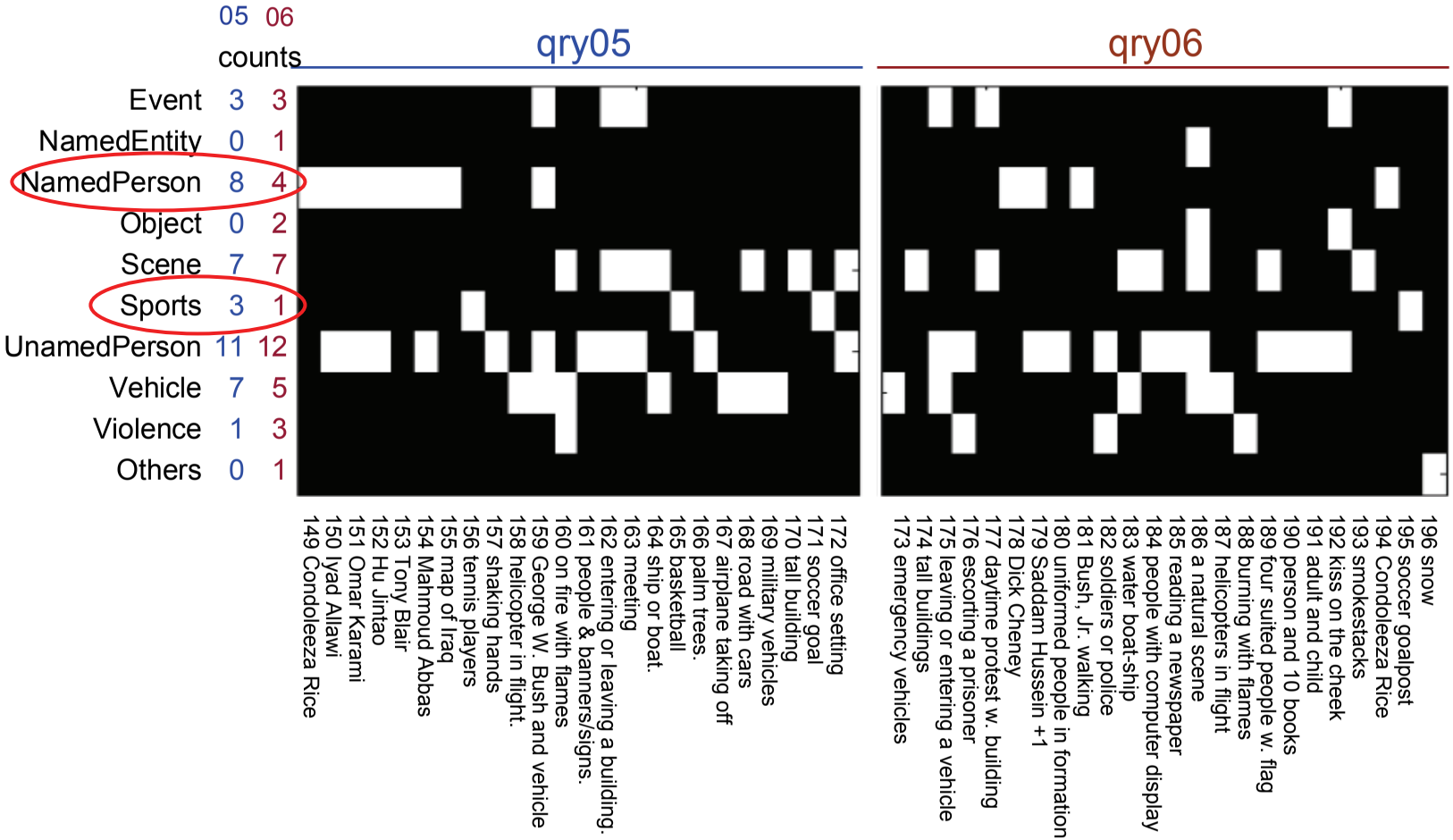
Named Person, Unnamed Person, Sports, Others

(b) 10 soft query components:

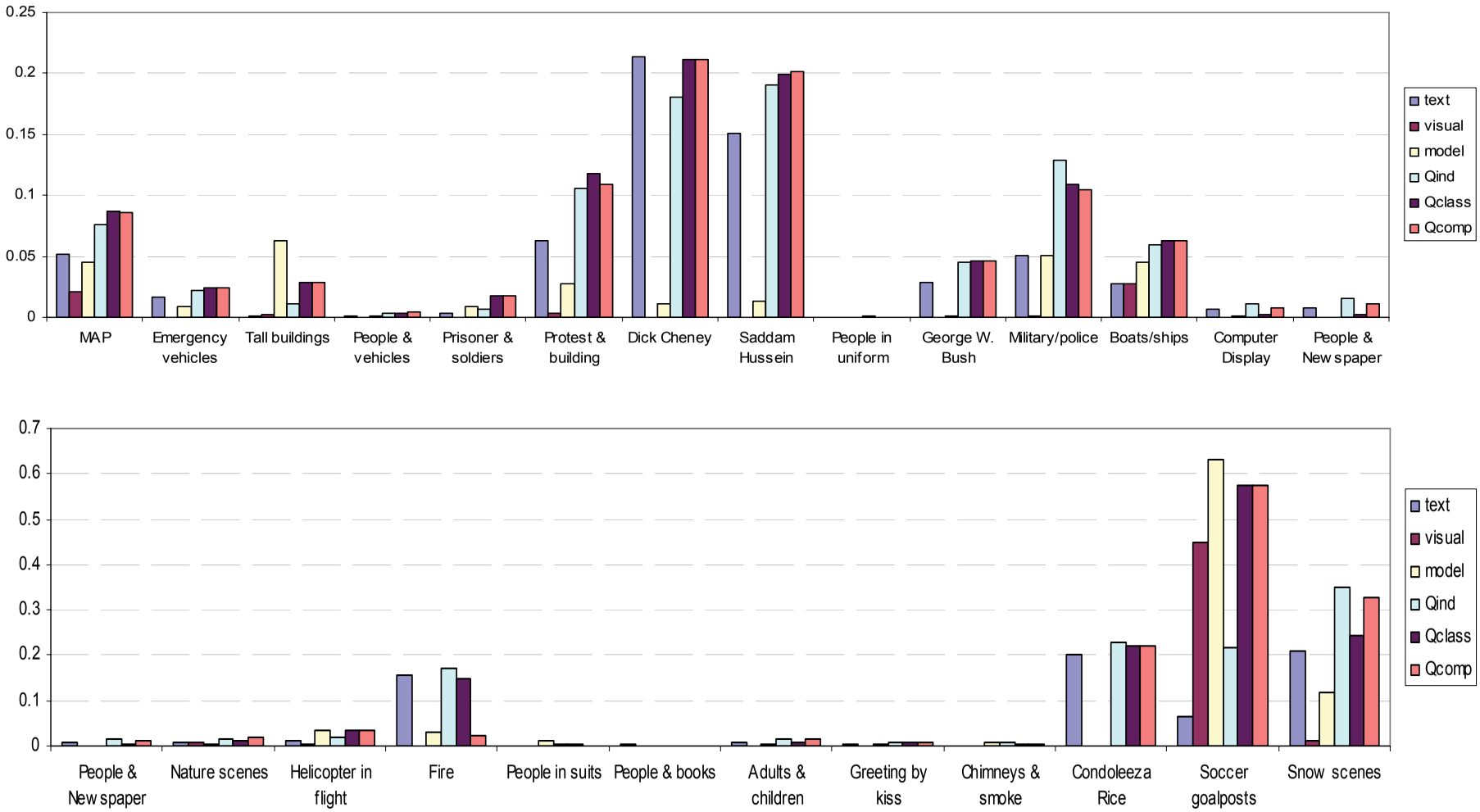
Named Person, Unamed Person, Sports, Named Entity, Event, Scene, Object, Vehicle, Weapon , Others



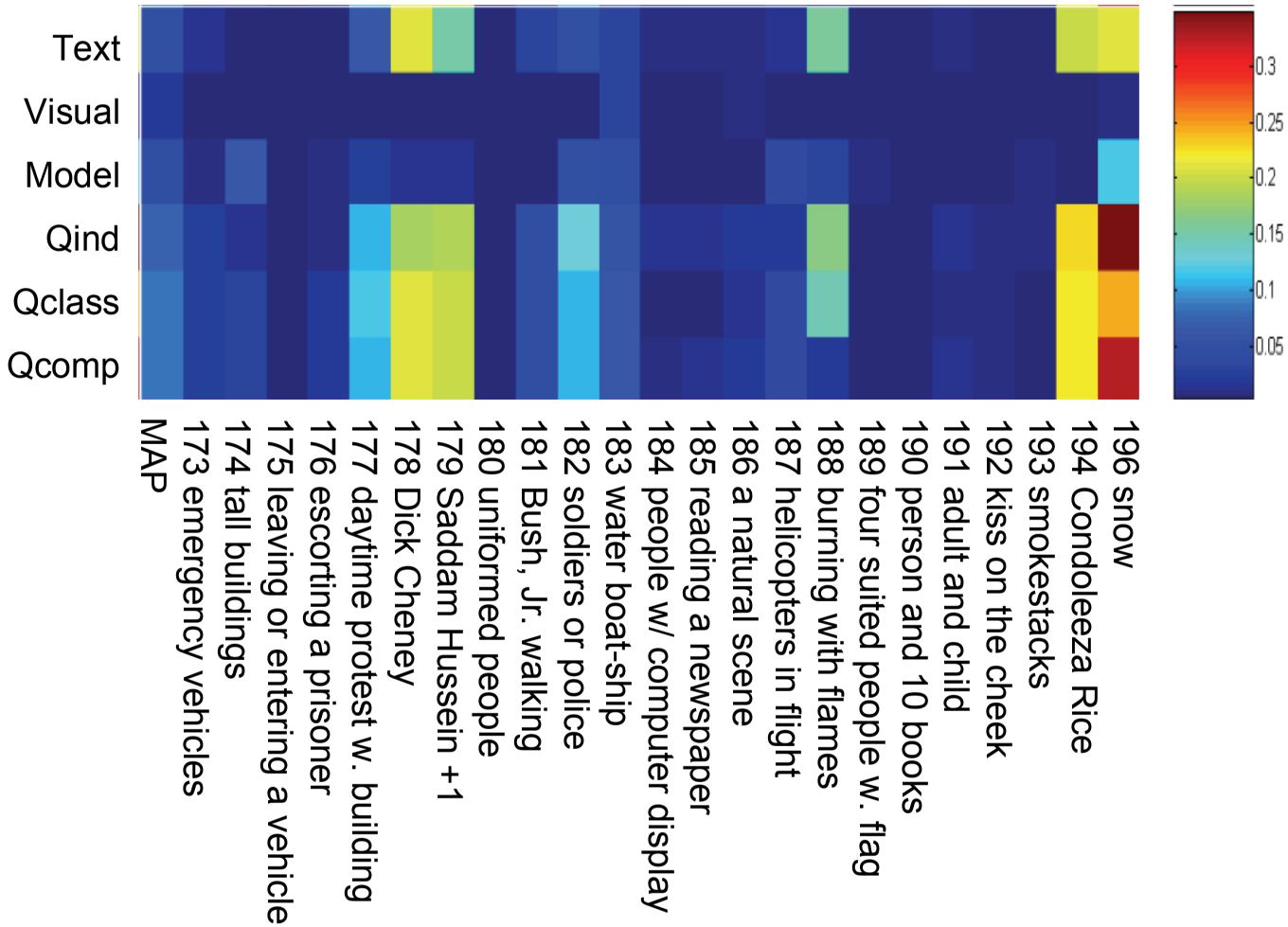
Query Components Coverage (2005-06 topics)



Multi-modal Fusion Results

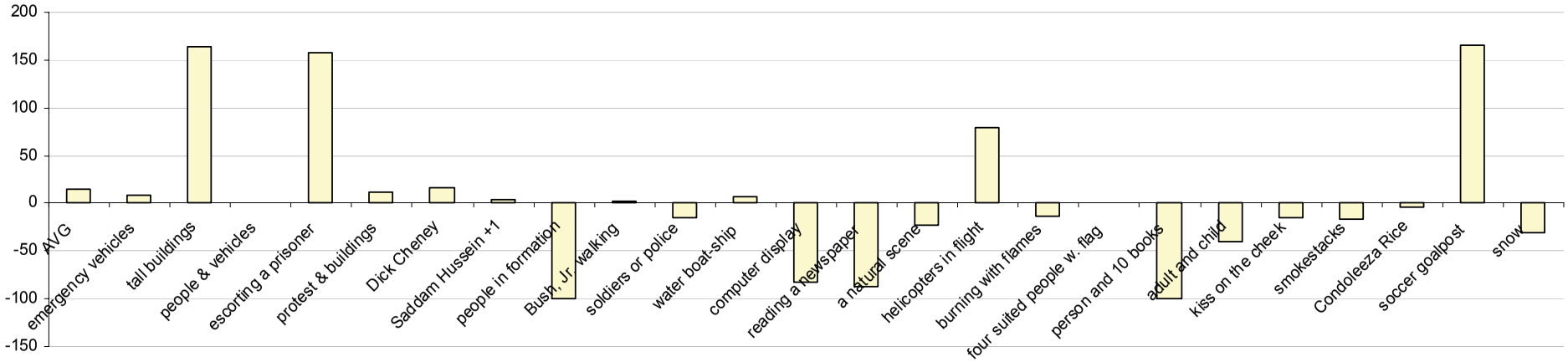


Multi-modal Fusion Results (Color-coded Performance)

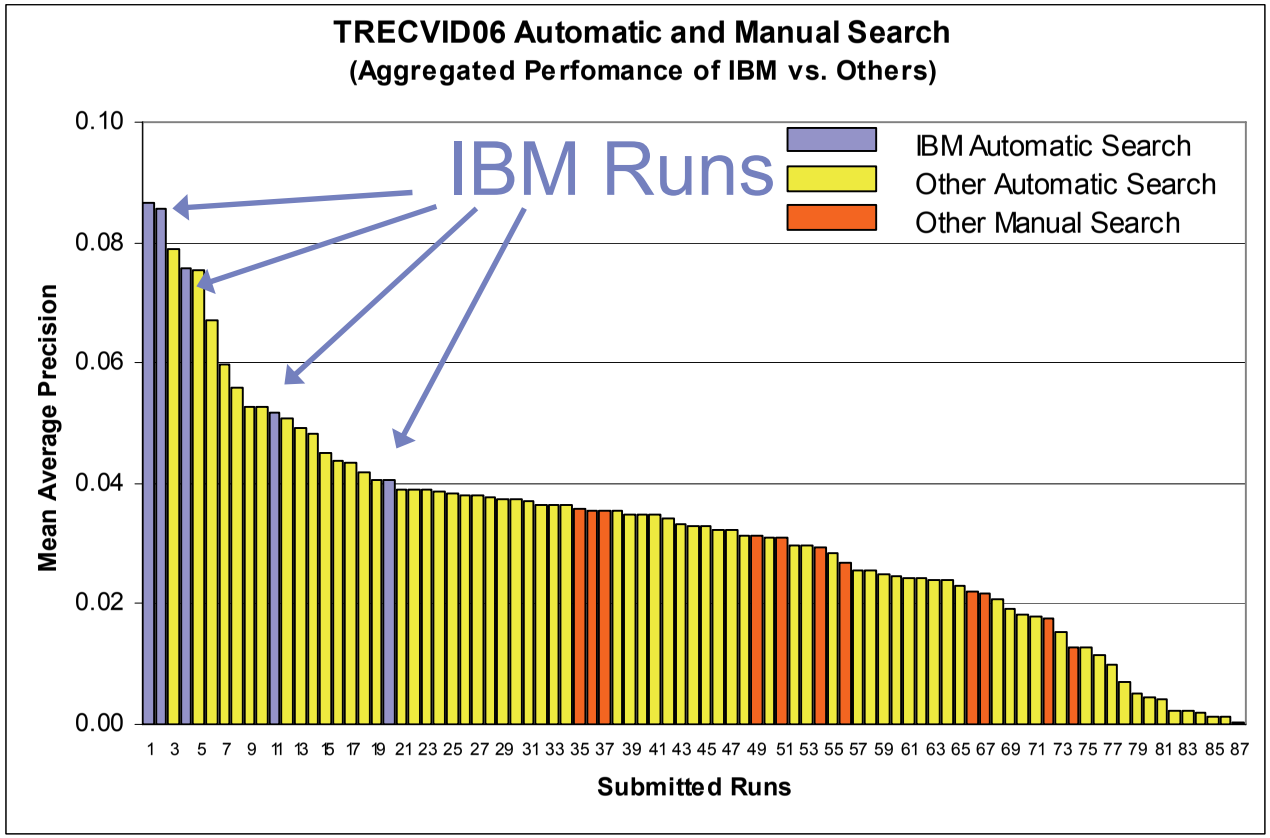


Multi-modal Fusion Results (Relative Performance)

- Relative improvement (%)
  - query-class fusion vs. query-independent fusion
- Observations
  - Concept-related queries improved the most:
    "tall building", "prisoner", "helicopters in flight", "soccer"
  - Named-entity queries improved slightly:
    "Dick Cheney", "Saddam Hussein", "Bush Jr."
  - Generic people category deteriorated the most:
    "people in formation", "at computer display", "w/ newspaper", "w/ books"



Automatic/Manual Search Overall Performance (Mean AP)



IBM Official Runs:

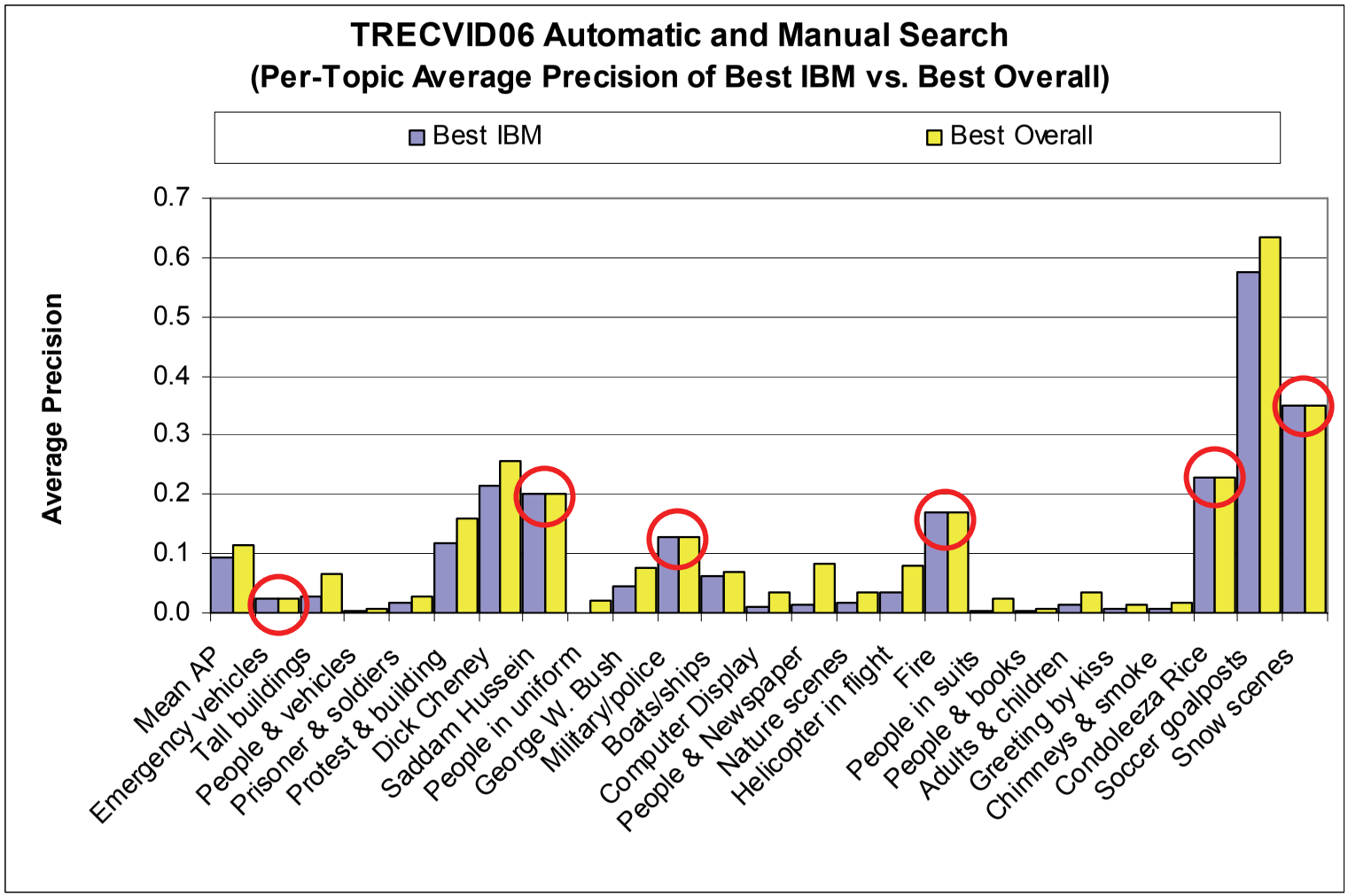
Text (baseline):	**0.041**
Text (story-based):	**0.052**
Multimodal Fusion:	
Query independent:	**0.076**
Query classes (soft):	**0.086**
Query classes (hard):	**0.087**

IBM Constituent Runs:

Model-based Run:	**0.045**
Visual-based Run:	**0.021**

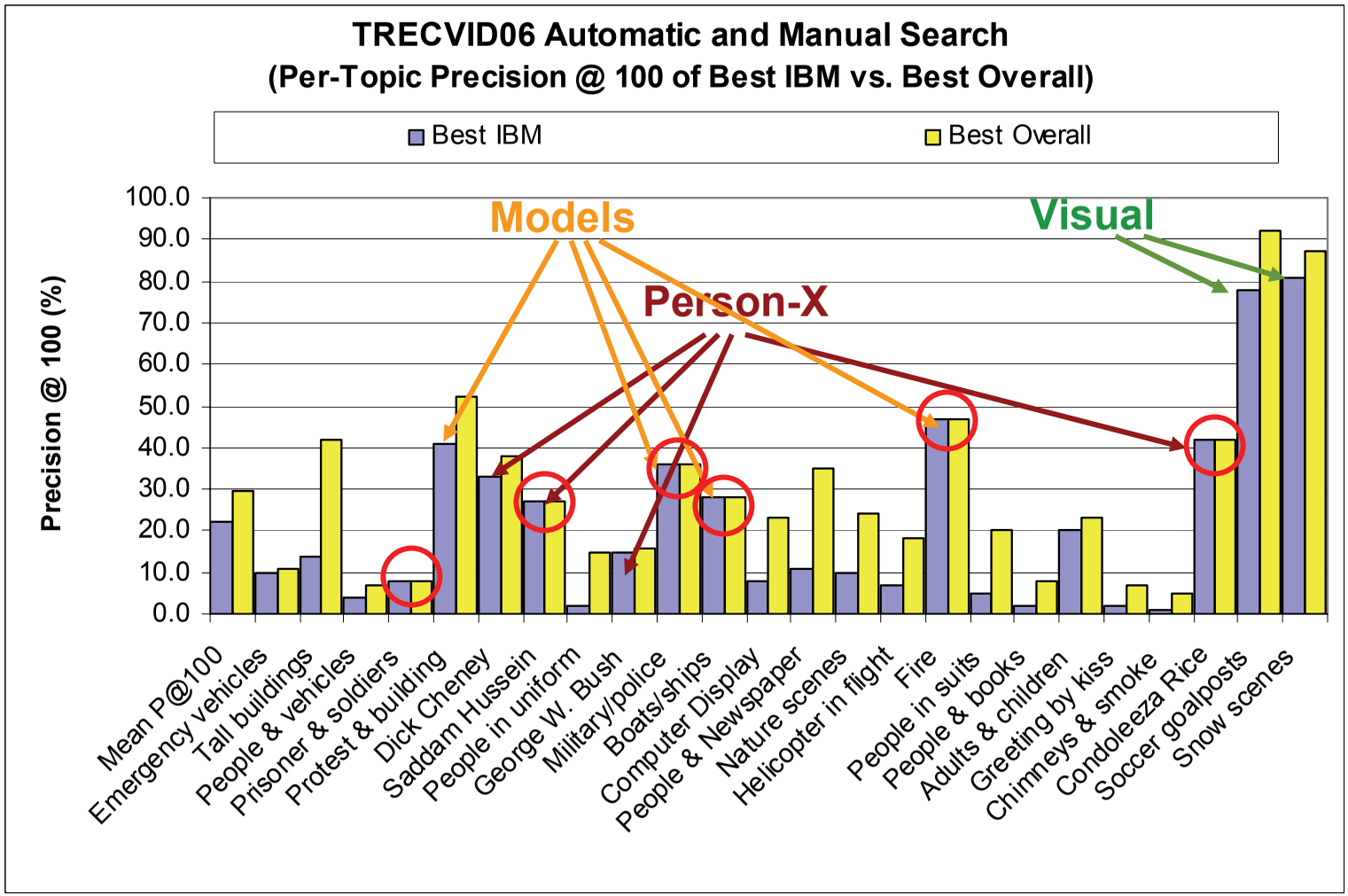
- Multi-modal fusion doubles baseline performance!
- Model-based retrieval and re-ranking instrumental in performance gain
- Visual retrieval not much of a factor this year

IBM vs. Others Per-Topic Analysis (Average Precision)



- Top performance on 6 out of the 24 query topics

IBM vs. Others Per-Topic Analysis (Precision at 100)



- Top performance on 6 out of the 24 query topics

Observations (~~2005~~ 2006)

- Visual retrieval ~~2x better~~ 3x worse than speech retrieval this year
  - Due to fewer sports topics and fewer near-duplicates
- Concept models helped significantly (~~>50%~~ 100% gain over baseline)
  - No domination by any single query class (e.g., Person-X, Sports, etc.)
- Automatic search ~~on par with~~ better than manual search!
  - Due to very few manual submissions
- **Search systems need to be comprehensive to be effective!!!**
- Proposals
  - Consider increasing # topics to reduce skew on aggregate measures
  - Consider standardizing query classes (Person-X, Sports, etc.)