Semantic Model-based Retrieval and Re-ranking

Problem

Approach

Semantic Concept Lexicon

- 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

Query Analysis

- Observations
  - Isolated words have ambiguous meaning
  - Words in vicinity provide context
- Phrase resolution
  - Provides sense disambiguation
  - Phrase = longest group of words with known meaning
- Dictionary: WordNet

Query Refinement

- Semantic Relatedness
  - WordNet Leak 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

Automatic Query Refinement

- Problem
  - Given arbitrary text query, automatically suggest additional relevant query keywords or semantic concepts related to the query topic
- Motivation
  - Improve recall by broadening narrow topics (too few hits)
  - Improve precision by disambiguating vague topics (too many hits)
- Approach
  - Automatically leverage large base of semantic models as query filters

Evaluation

TRECVID 2005 corpus

- 160 hrs. of U.S., Arabic, and Chinese news
- Speech transcripts based on
- Automatic Speech Recognition
- Machine Translation (as needed)

TRECVID 2005 query topics

- Named people (Person-X)
- Unnamed people
- Events/Actions
- Objects
- Scenes/settings

Training

- Models parameters tuned on TRECVID 2005 development set
- Search parameters tuned on TRECVID 2003 corpus and topics
- Applied to TRECVID 2005 test set

Approach Overview

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 Leak similarity
3. Query execution
   - Semantic-based retrieval using statistical concept models
   - Weighted averaging model fusion

Evaluation:

- Fusion outperforms isolated runs: 50% overall improvement
- Improvement on named entities: 23%
- Improvement on unnamed entities: 89%