ECE8813
Statistical Natural Language Processing

Lecture 6: Class Project List

Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
chl@ece.gatech.edu

Project List

1. Word sequence modeling and applications
2. Text categorization, topic identification and tracking
3. Information retrieval: document indexing, retrieval, search engine
4. Web page ranking, clustering, and classification
5. Message understanding: e.g. spam mail classification
6. Part-of-speech tagging and sentence structure parsing
7. Automatic image annotation
8. Automatic recognition of speaker, speech and language
9. Speaker segmentation in audio and video
10. Voice and face morphing: voice and image quality manipulation
11. Voice annotation and retrieval of photos
12. Video shot segmentation, classification, clustering
13. Image classification: e.g. spam image classification
14. Classification of genre, instrument, singer in music
15. Audiovisual event detection in video: e.g. clap, anchor, scoring
16. Financial data analysis: regression, classification and prediction
17. Bioinformatics: plenty of data out there
18. Design your own learning applications: bring your own data sets
19. Any others? Propose a team project if it justifies the effort

Project Report

- Introduction: literature survey
  - Problem definition and potential applications
- Problem formulation
  - Chosen approaches: new or existing from ECE7252 topics
  - Preliminary findings, if any, on a small pilot dataset
- Experimental configurations and results
  - Training, validation and test corpus
  - Implementation issues: tools, codes and demo
  - Evaluation metric
  - Tabulation and plotting of experimental results
  - Qualitative and quantitative analysis
- Concluding remarks: findings, difficulties and summaries
- References

Project Planning

- Expected effort: 4-5 weeks, +30% of your grade !!
  - Pick a ECE7252 subject, define your project and have fun !!
- Supporting materials
  - Data set for training, validation and test
  - Literature survey, tool and code availability
- Designing and planning process
  - Write a project proposal: laying out problem definition, approaches, supporting tools, evaluation metrics, estimated level of effort
  - Submit the proposal and we will iterate
  - Agree to a proposal by the end of February (talk to me !!)
- Execution: time management is job 1, start now !!
- Report and presentation:
  - Report due before the final week (we have no Final Exam)
  - Presentation during the last two weeks (15 minutes each)
- Consultation: Talk to me before you invest major effort, you want to finish the project, not to leave it half done !!

N-gram Modeling

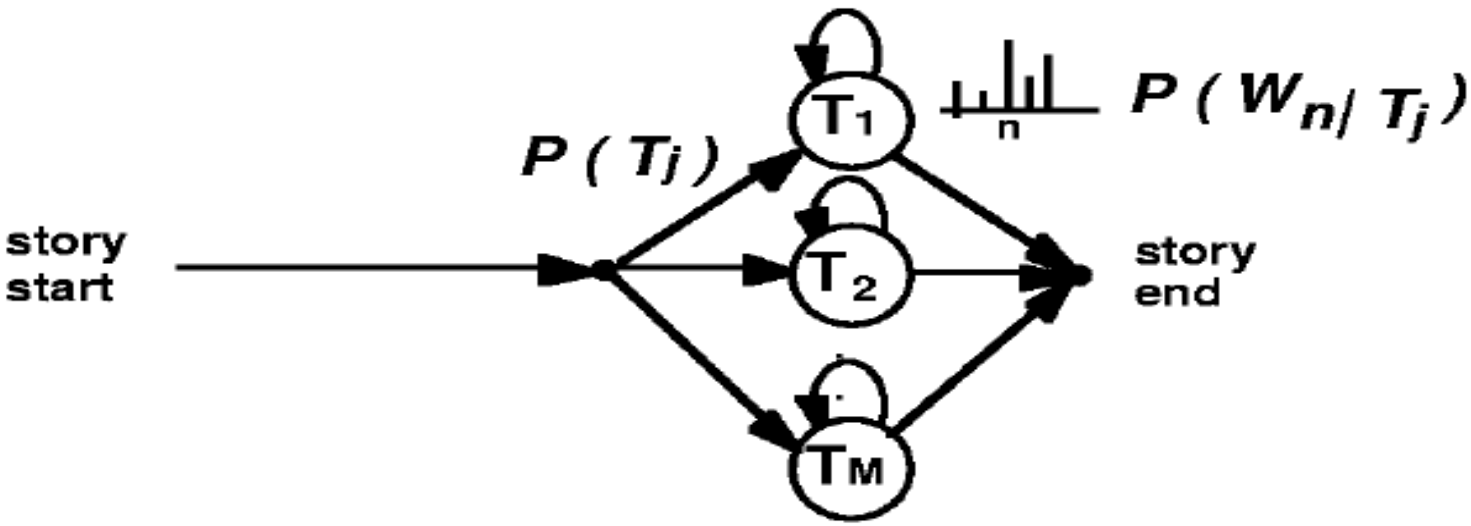
- Domain-specific sentence modeling
  - Purpose: build *n*-grams and use them to rank sentences
  - Training Corpus: 1.5 million WSJ sentences
  - Testing Corpus: unseen WSJ sentences
  - Techniques involved:
    1. N-gram modeling
    2. Computing sentence probability
    3. Bag-of-word modeling
    4. DP Viterbi search: finding the most likely word sequence
- Domain-specific term clustering
  - Doing the same but finding words belonging to a group

Text Categorization (TC)

- Domain-specific topic modeling and classification, TC is also known as topic identification
  - Purpose: building topic models to classify unseen documents
  - Training Corpus: 7000 documents from Reuter with topic tags
  - Testing Corpus: unseen Reuter documents with topic tags
  - Techniques involved:
    1. Vector-based document representation
    2. Latent semantic indexing based feature extraction
    3. Vector-based distance measures, and scoring
    4. Evaluation metric: precision, recall, and F1 measures
    5. Topic classifier design: vector-based classification algorithms, e.g. SVM, LDF, and others
- Other related problems

Topic Tracking: Decoding

- Put all topics in one network (like isolated-word ASR)
- Viterbi search → optimal path → recognized topic
- Each state is attached with an n-gram model, which is estimated from all training documents of that topic



Information Retrieval (IR)

- Document indexing and retrieval
  - Purpose: building a search engine to index and retrieve text documents (Google-like keyword based search)
  - Training Corpus: 7000 documents from Reuter
  - Techniques involved:
    1. Term-document matrix (also known as a routing matrix) building
    2. Latent semantic indexing based document representation
    3. Vector-based distance measures, and scoring
    4. Evaluation metric: precision, recall, and F1 measures, efficiency
- Other related problems

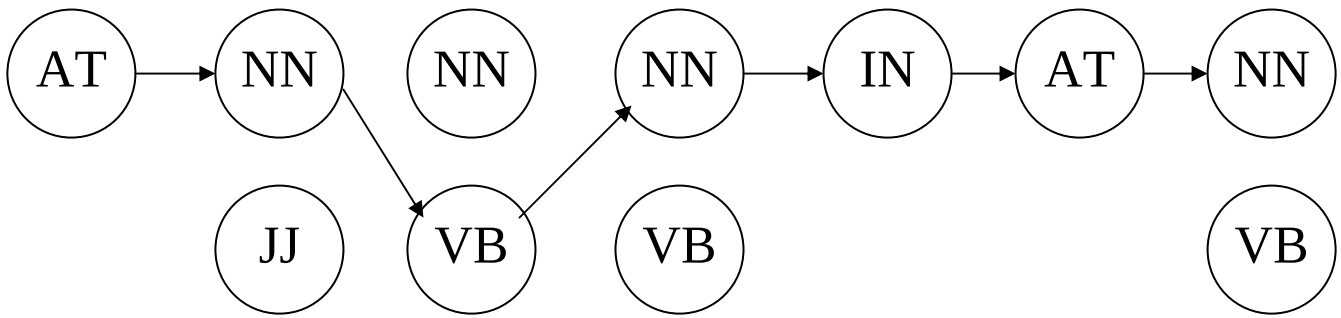
Part-of-Speech Tagging

- For English
  - Purpose: building a PoS tagging system to assign a sequence of PoS tags to an unseen sentence
  - Training Corpus: WSJ but tags are needed
  - Testing Corpus: WSJ
  - Techniques involved:
    1. Assigning initial tags to a small set of sentences, and bootstrapping to a larger set
    2. N-gram modeling of tag language models
    3. N-gram modeling of tag-specific language models
    4. Viterbi decoding of the most likely tag sequence

Part-of-Speech (POS) Tagging

- Finite state network (FSN) representation
  - State (node) space: the set of tags
  - Arc: tag transition (probabilities)
  - State output: tag-specific word probabilities
  - State-sequence: tag sequence
- An example:

The representative put chairs on the table.



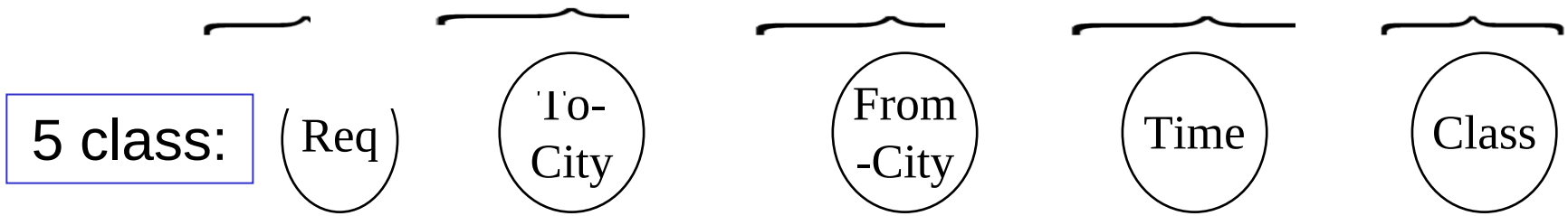
Message Understanding

- Domain-specific text understanding
  - Purpose: building a concept decoding system to assign a sequence of concept to an unseen sentence so that messages behind the sentence can be decoded
  - Training Corpus: ATIS but tags are needed
  - Testing Corpus: Airline Travel Information System
  - Techniques involved:
    1. Assigning initial concept tags to a small set of sentences, and bootstrapping to a larger set
    2. N-gram modeling of concept language models
    3. N-gram modeling of concept-specific language models
    4. Viterbi decoding of the most likely tag sequence

Concept Understanding

- Finite state network (FSN) representation
  - State (node) space: the set of concepts
  - Arc: concept transition (probabilities)
  - State output: concept-specific word sequences
  - State-sequence: concept sequence (meaning expressed in sequence of semantic attributes)
- An example:

I want to fly to Boston from Dallas Friday noon on coach.

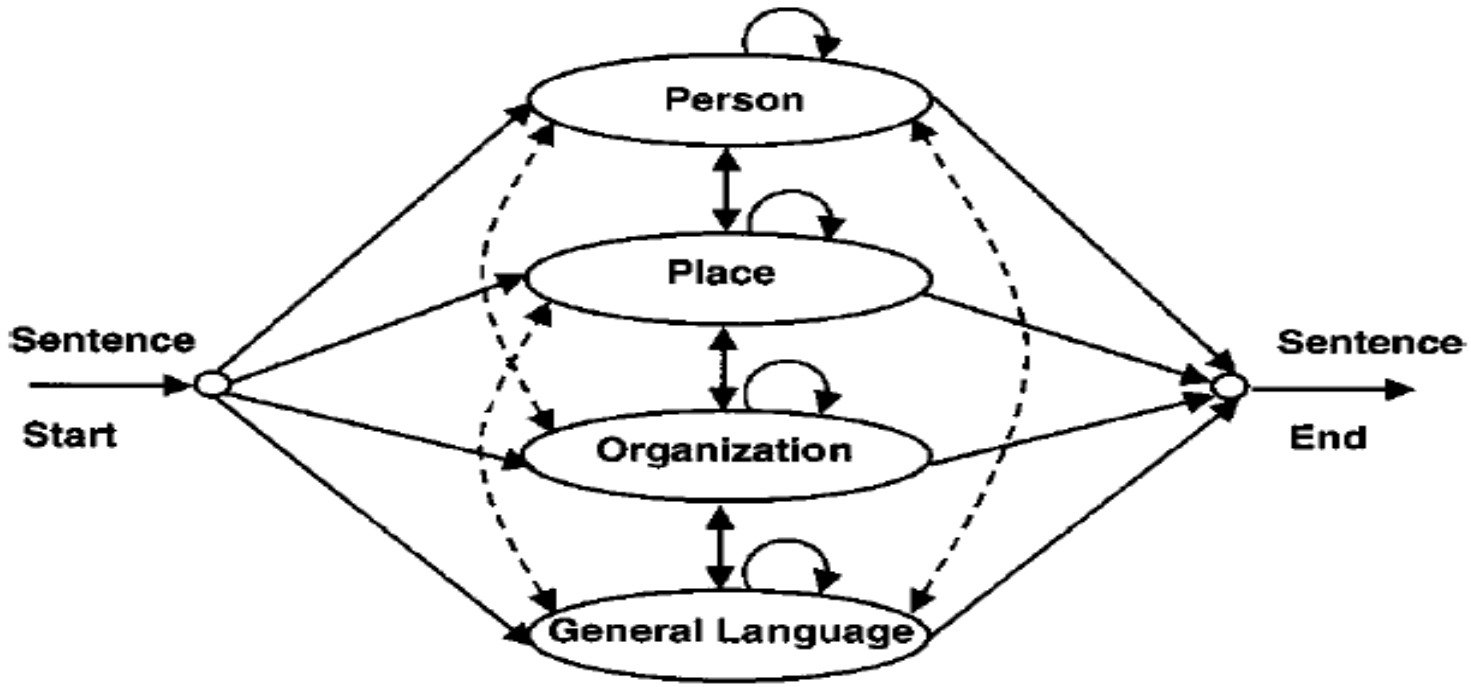


Name Spotting: A Similar Application

- Name spotting is doable and a useful task
- Explicitly model all possible name classes:
  - Person's names
  - Organizations
  - Locations
  - Dates
  - Times
  - Numerical expressions: money, percent
  - NOT-A-NAME: other general language parts
- Each class is modeled as a bigram-like statistical model

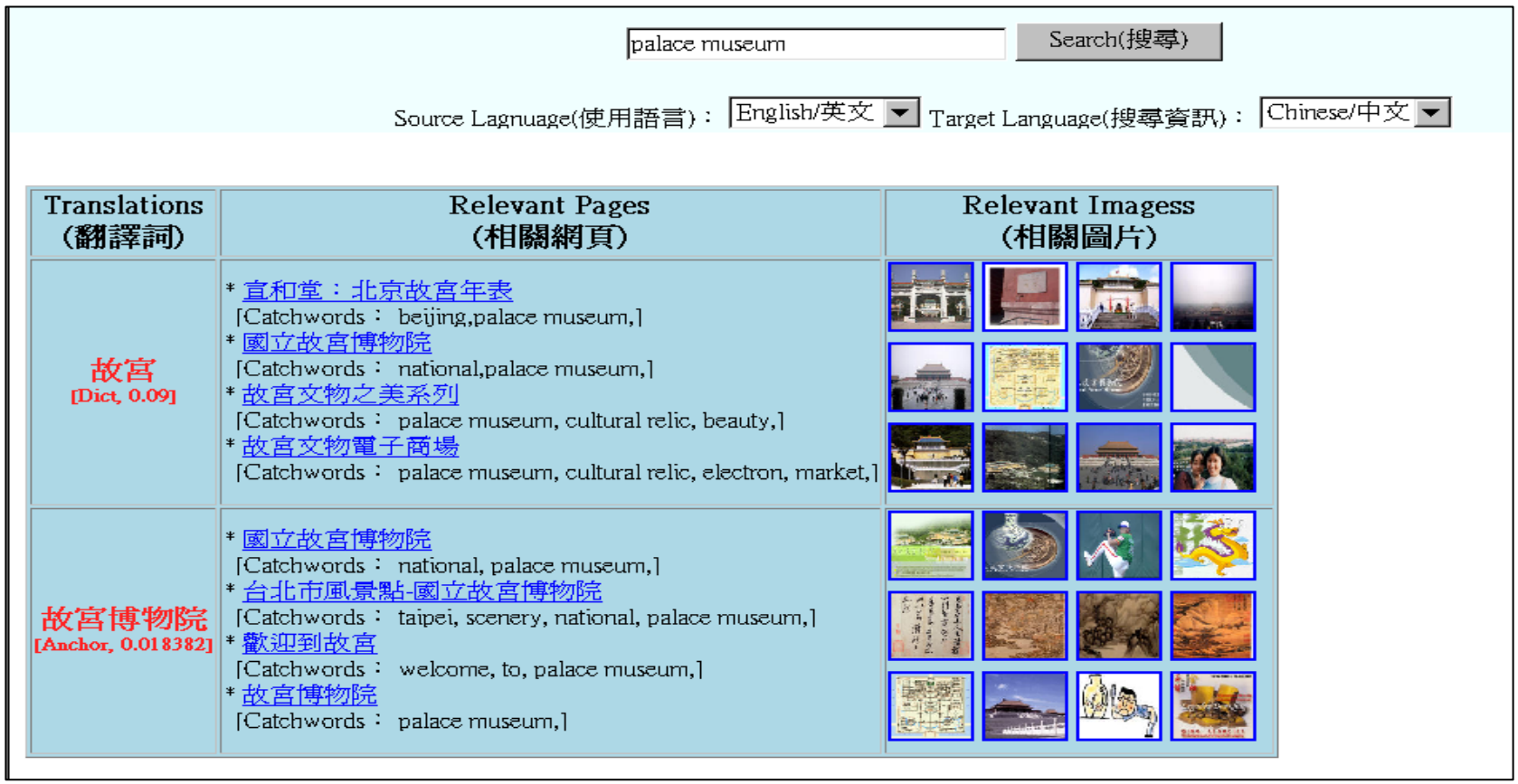
Name Spotting: Decoding

- Put all together to build a search network
- Viterbi search → backtrack the optimal pass → optimal name-class labels



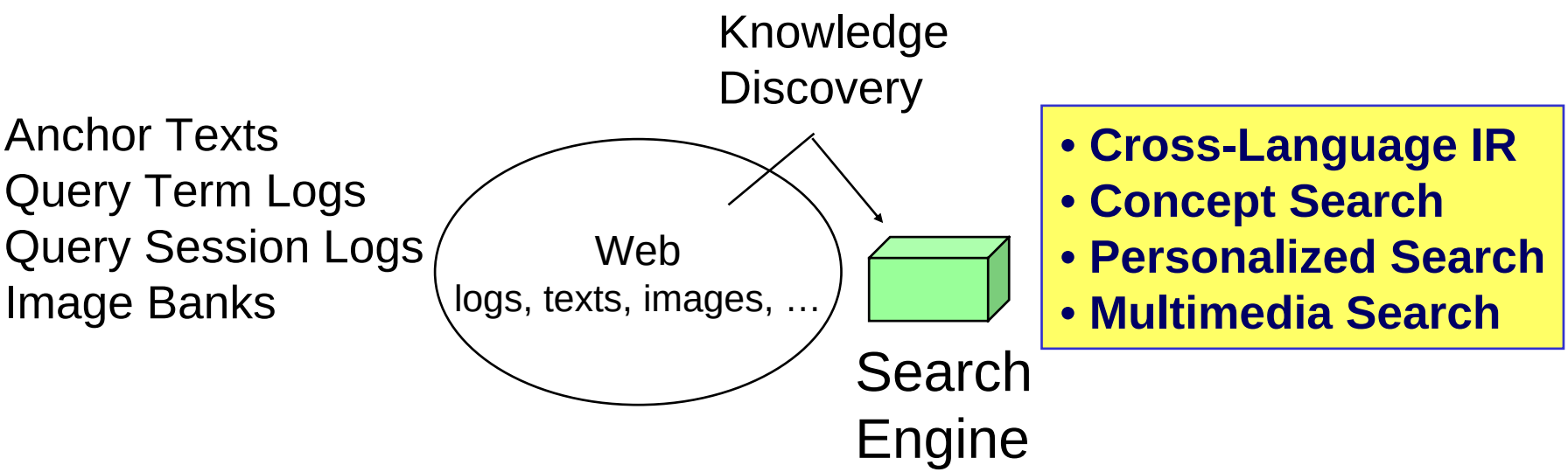
Cross-Language Web Search (IIS/Taiwan)

- Allows users to query in one and search for pages and documents that are written or indexed in another language

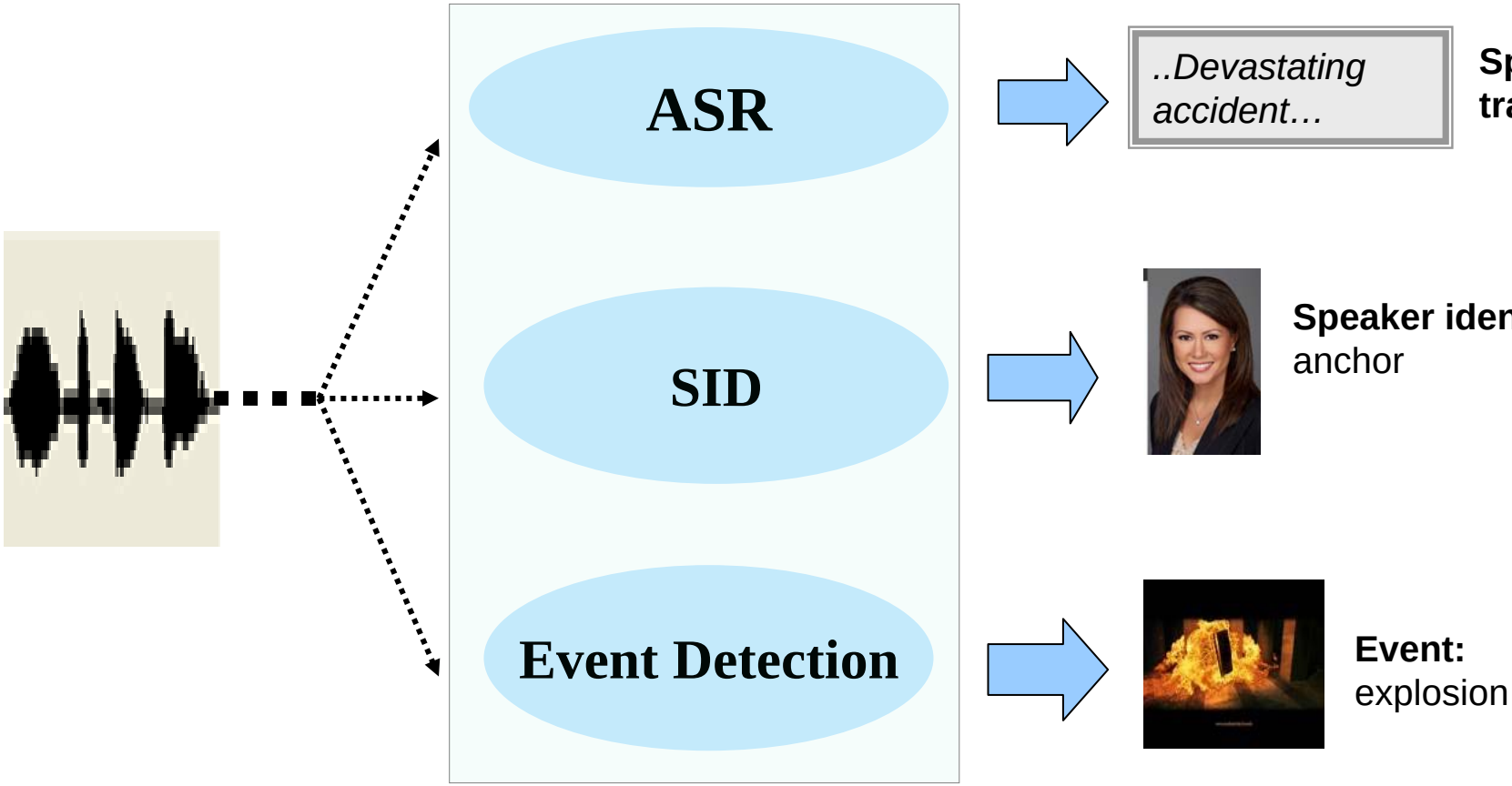


From Web Search to Web Mining

Exploring the Development of Advanced IR Techniques through Web Mining

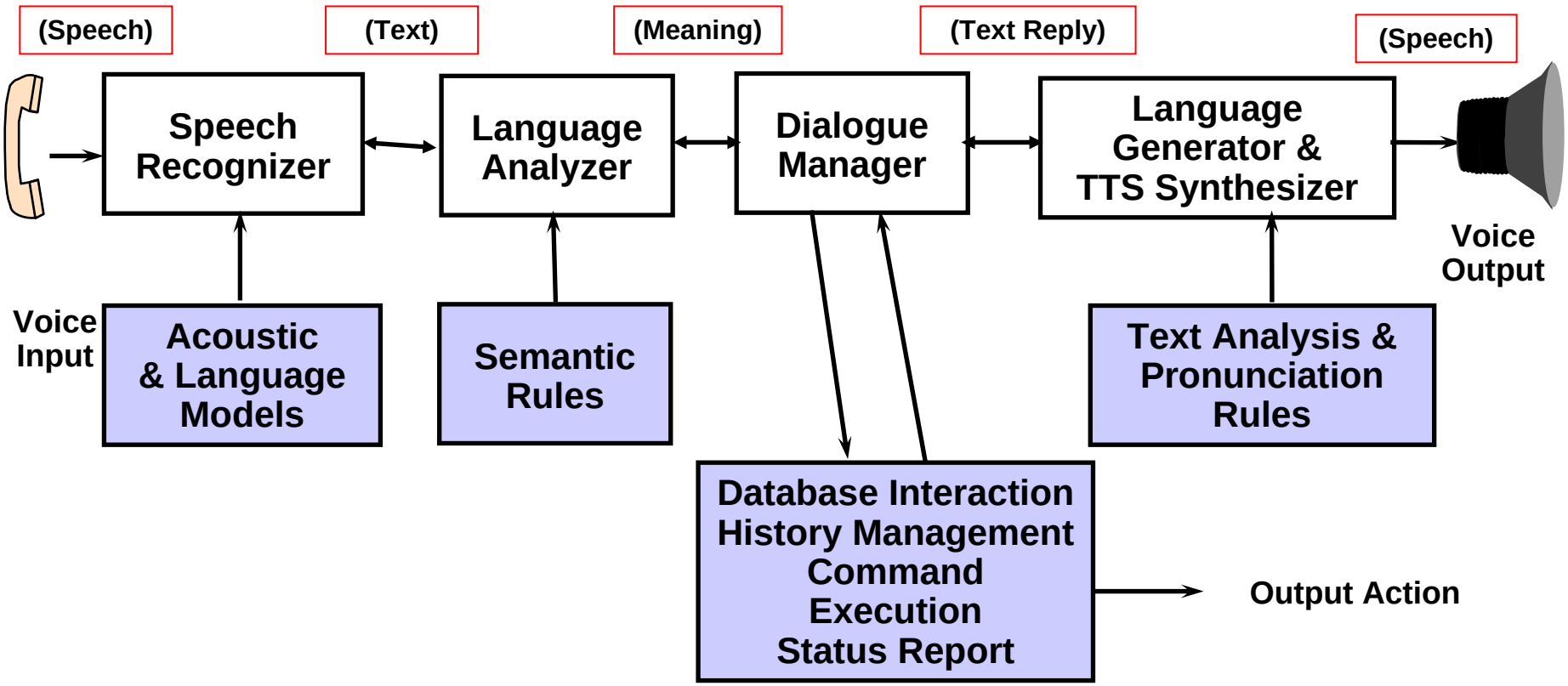


Speech and Speaker Data Mining

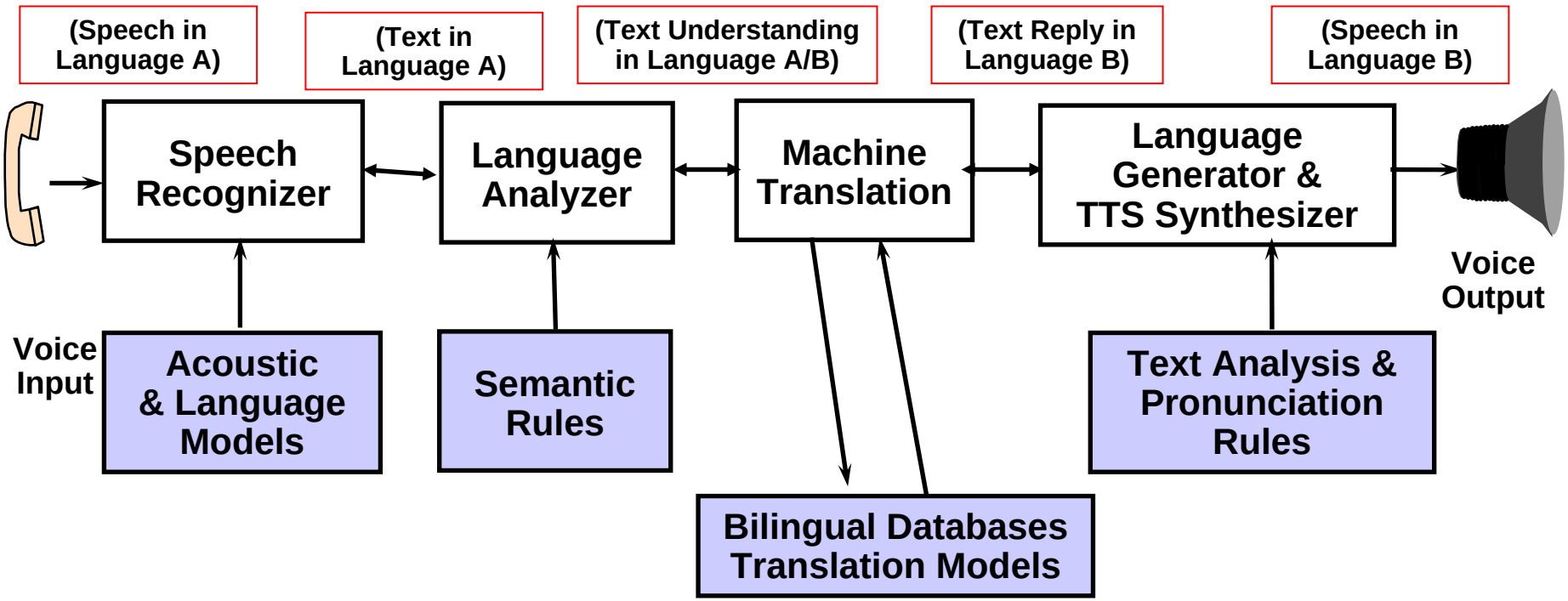


Automatic Speech Recognition (ASR)
Automatic Speaker Identification (SID)

Conversational User Interface – R2D2



Universal Speech Translation – C3PO



Spoken Language Identification (LID)

- Following Shannon's study of English
  - Purpose: building a system to identify the language corresponding to a spoken utterance
  - Training Corpus: OGI six-language corpora
  - Testing Corpus: similar corpora
  - Techniques involved:
    1. Finding acoustic alphabets and building corresponding models
    2. Tokenizing utterances into acoustic alphabet sequences
    3. Converting each utterance into a spoken document vector
    4. Building vector-based language classifiers
    5. Performing spoken language identification

Music Genre Classification

- Following Shannon's study on English letters
  - Purpose: building a system to identify the music style corresponding to a audio passage
  - Training Corpus: TBD
  - Testing Corpus: TBD
  - Techniques involved:
    1. Finding audio alphabets and building corresponding models
    2. Tokenizing music passages into audio alphabet sequences
    3. Converting each passage into an audio document vector
    4. Building vector-based genre classifiers
    5. Performing music genre identification
- Similar problem: spoken language identification

Altavista Image Search

Query: Tiger

Tiger.jpg

No Caption

Ranking: 1

tiger.jpg

No caption

Ranking: 4

Tiger__1_11.jpeg

No Caption

Ranking:18

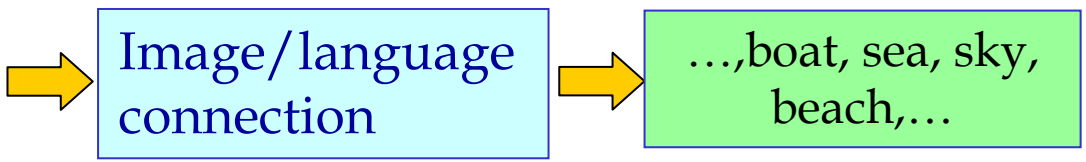
tiger-003.jpg

No Caption

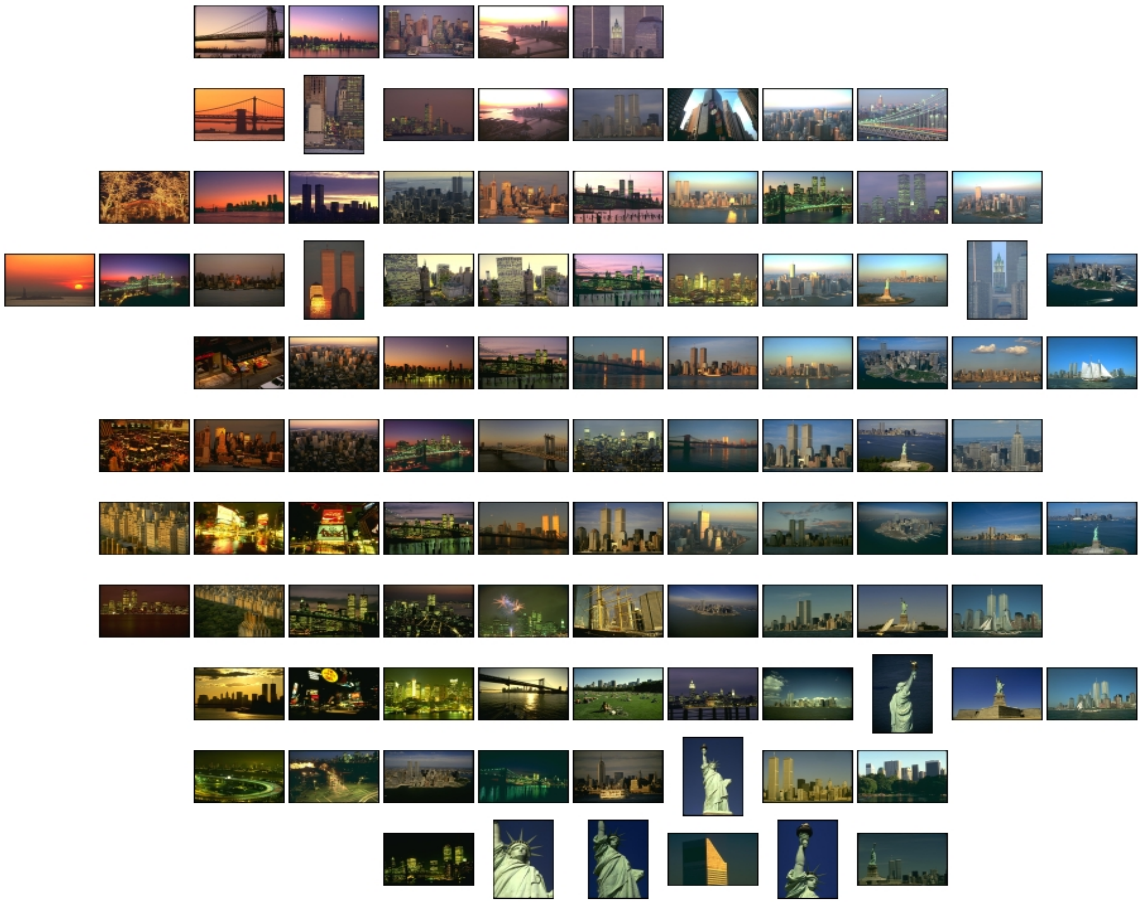
Ranking: 34

Automatic Image Annotation (AIA)

- Purpose: A process associating concepts or keywords to images describing their visual content
  - Training Corpus: 4500 images from Corel
  - Testing Corpus: 500 Corel images
  - Techniques involved:
    1. Extracting image features and producing visual alphabets
    2. Forming words and converting images into vectors
    3. Latent semantic indexing based feature extraction
    4. Multi-topic topic classifier design

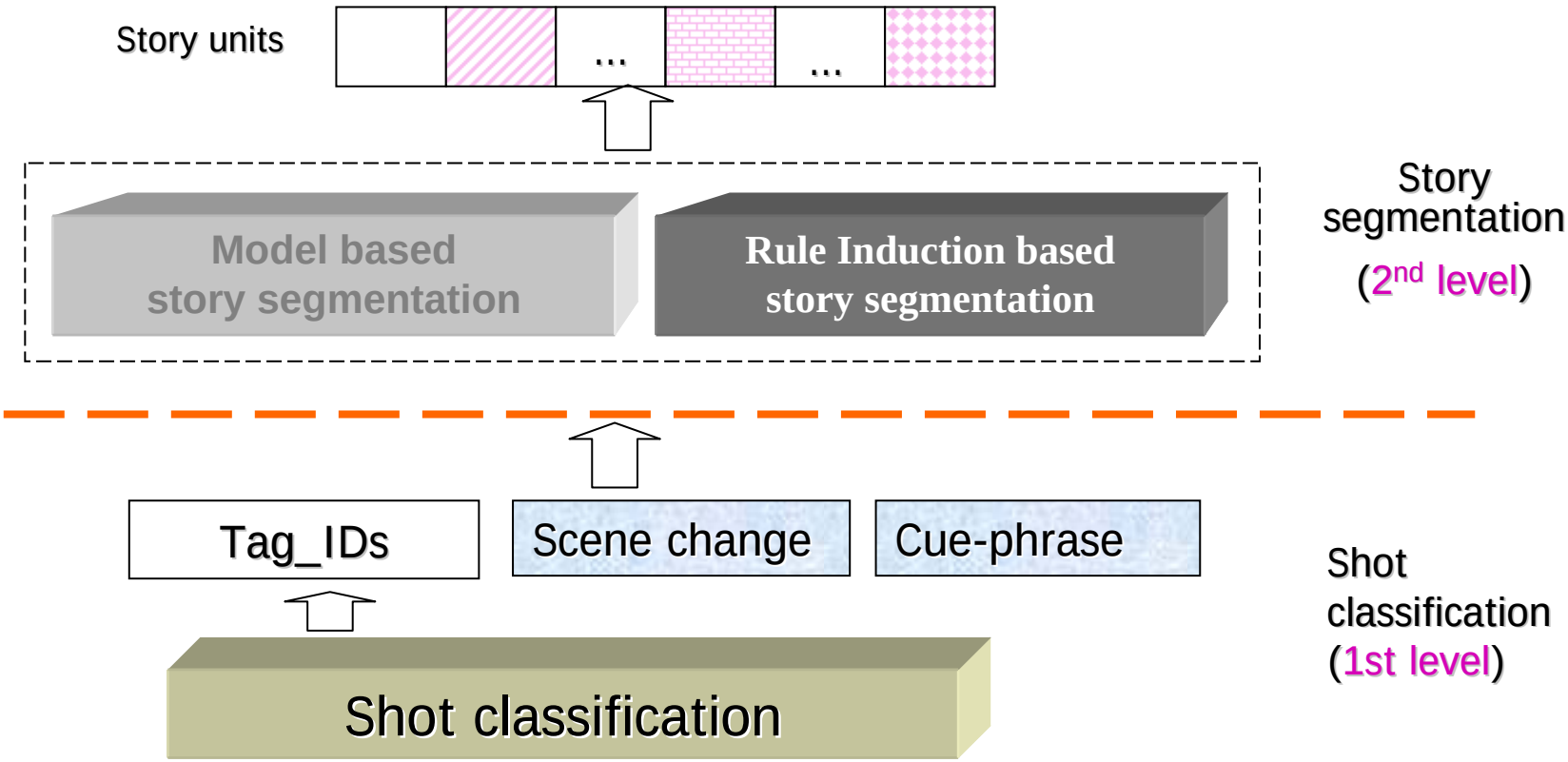


Voice and Text Based Photo Retrieval



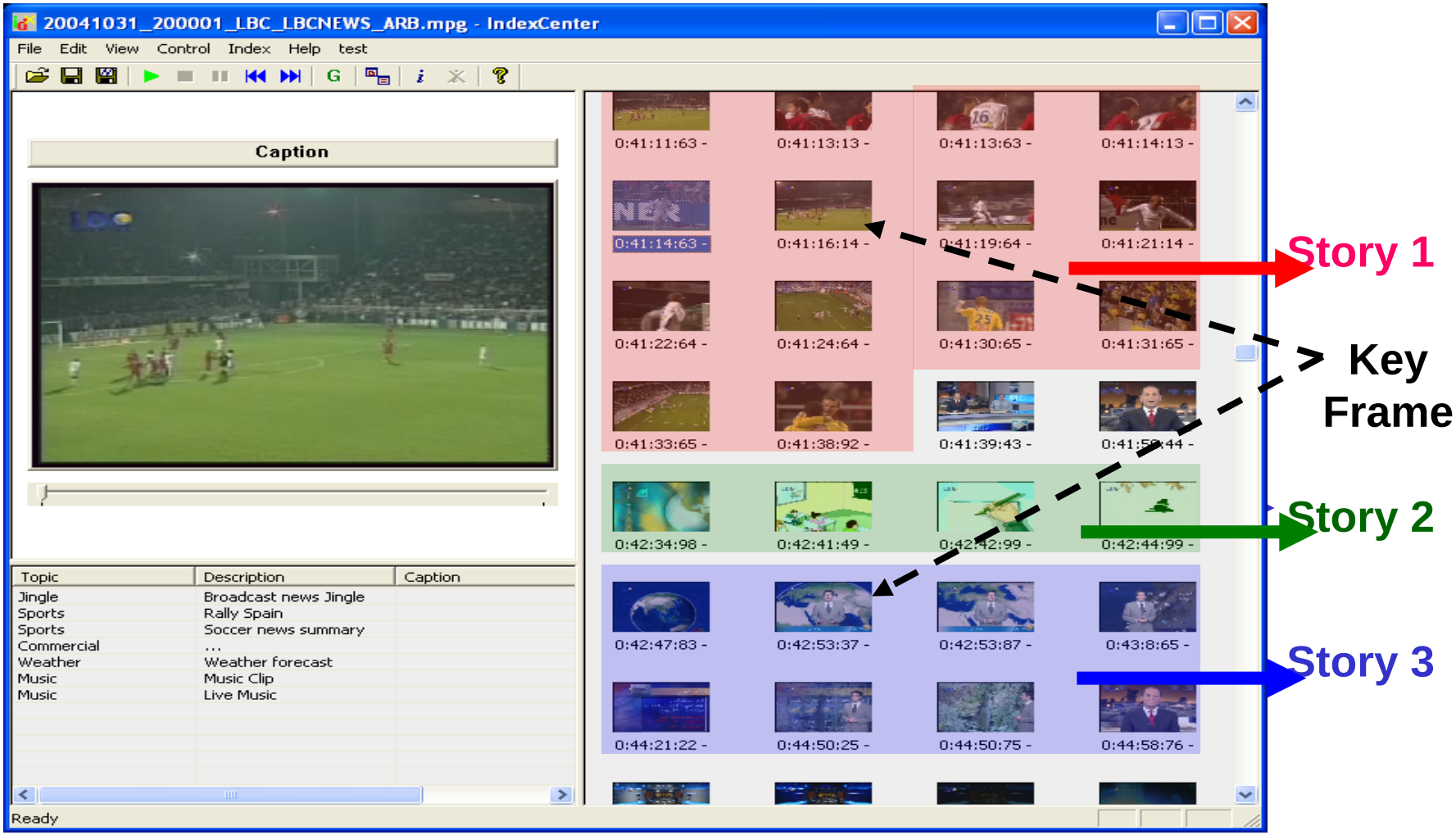
- Voice and Text annotation of photos
- Indexing and retrieval of photos
- Content based example search does not give good performance
- Concept based keyword search
  - GUI
  - Speech UI
  - Multimedia UI

Video Processing and Representation

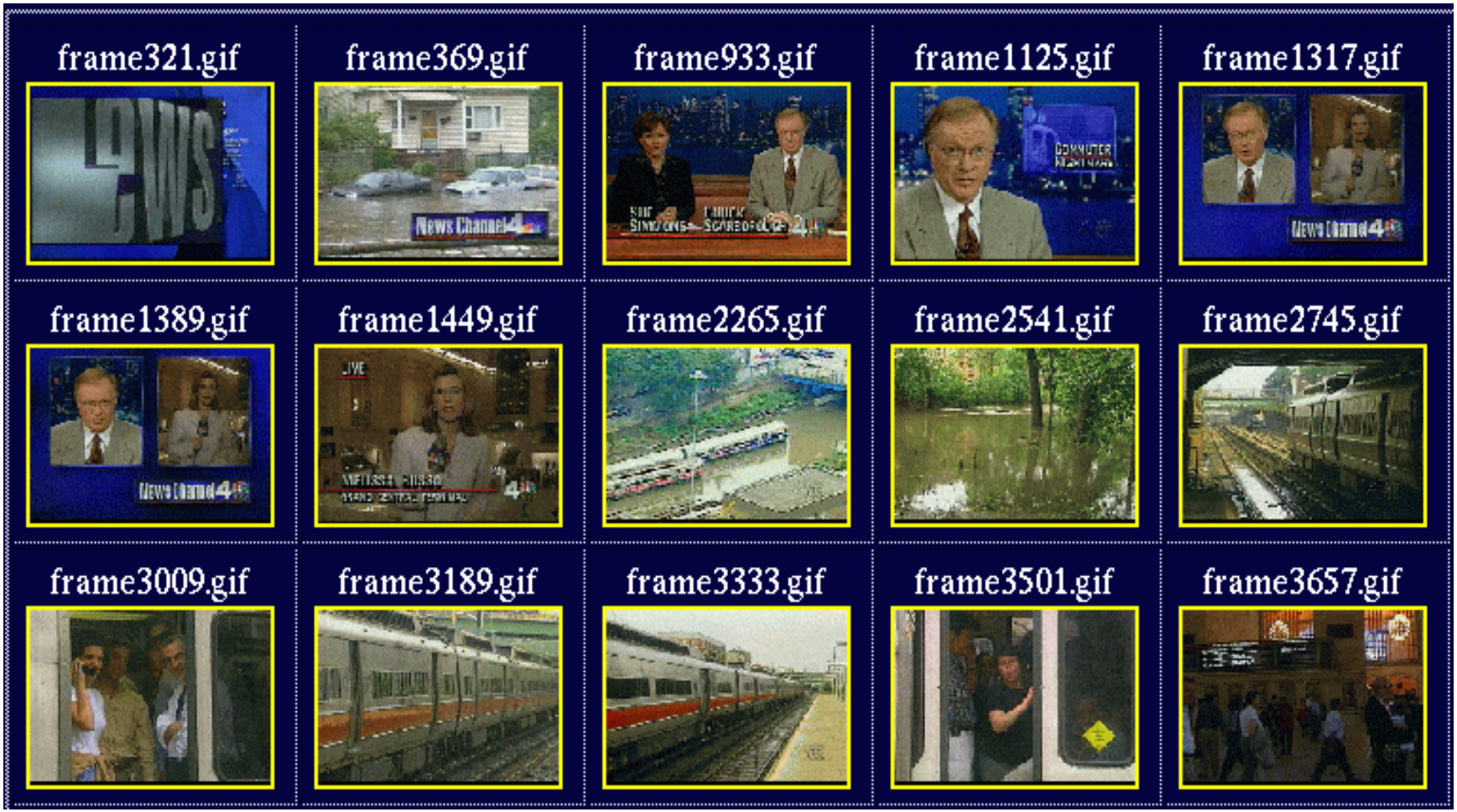


TRECVID is a community-supported annual open evaluation of technologies: for topic detection and tracking of multiple thread of similar stories spanning over a period of time, and from multiple channels, and covering multilingual sources

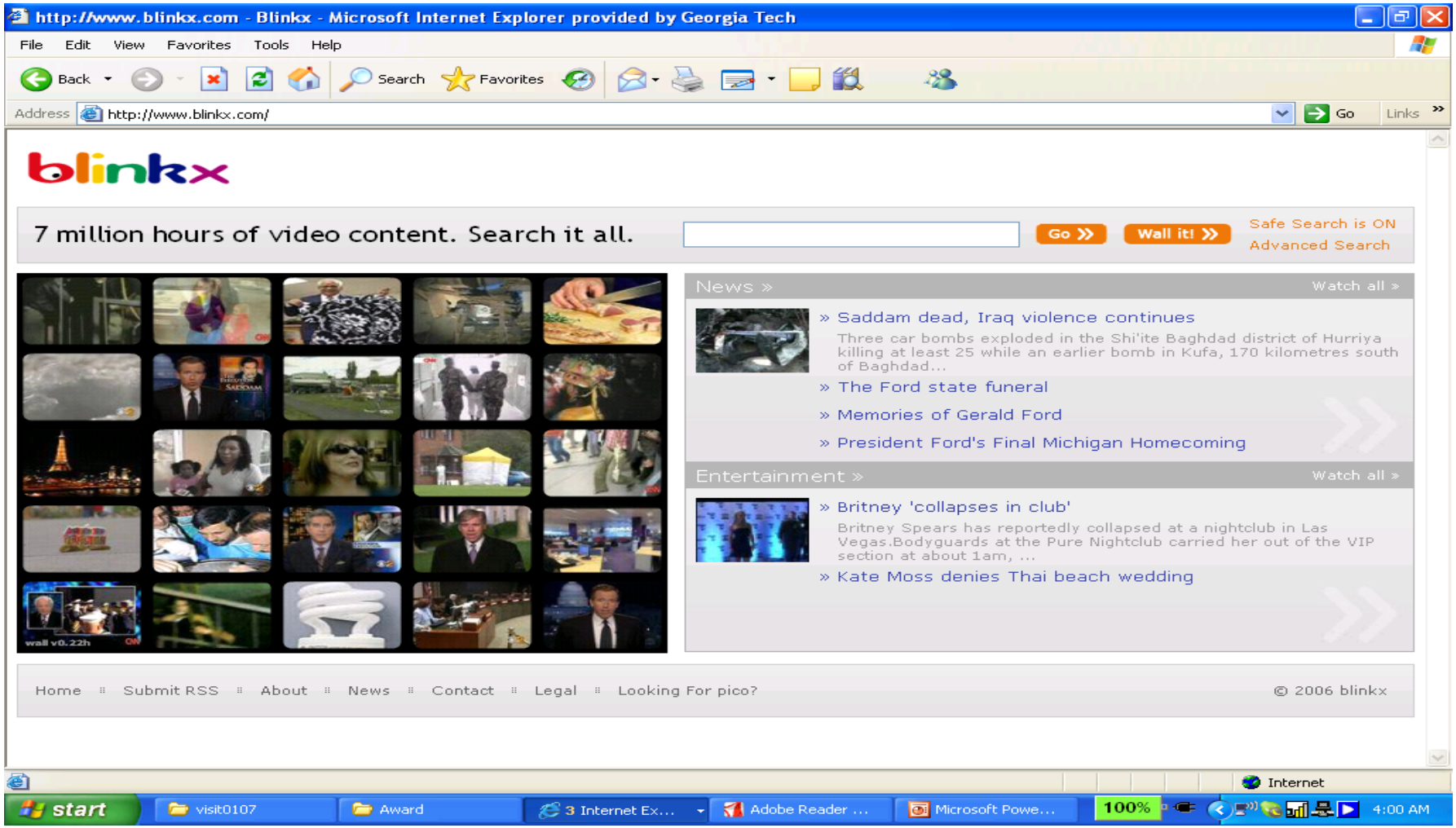
Video Shot Segmentation



Video & Audio Story Segmentation (1st Step to Indexing & Retrieval)



Blink-X: A Video Search Portal



Other Software Packages

- HTK: speech modeling kits (for hidden Markov model)
- GMTK: graphical model took kit (for speech/language)
- LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- NETLAB: http://www.ncrg.aston.ac.uk/netlab/
- CMU AI Repository
  - http://www.cs.cmu.edu/afs/cs/project/ai-repository/ai/areas/learning/systems/0.html
- JMLR machine learning open source software
  - http://jmlr.csail.mit.edu/mloss/
- Weka: data mining tool in Java
  - http://www.cs.waikato.ac.nz/ml/weka/
- R: http://www.r-project.org/
  - A free alternative to S-Plus developed at Bell Labs
  - If you know C, you will be right at home with R

Machine Learning Dataset Links

- UCI machine learning repository
  - http://archive.ics.uci.edu/ml/
- Open Directory Project:
  - http://www.dmoz.org/Computers/Artificial_Intelligence/Machine_Learning/Datasets/
- Datasets for knowledge discovery
  - http://www.kdnuggets.com/datasets/
- Machine learning & data mining: face, objects, etc.
  - http://cervisia.org/machine_learning_data.php
- BBC datasets: news and sports
  - http://mlg.ucd.ie/content/view/21/

References

1. S. Gao, H.-D. Wang and C.-H. Lee, “Automatic Image Annotation through Multi-Topic Text Categorization,” submitted to ICASSP2006, Toulluse, 2006.
2. B. Ma, H. Li and C.-H. Lee, “An Acoustic Segment Modeling Approach to Automatic Language Identification,” *Proc. InterSpeech-2005*, Lisbon, Portugal, September 2005.
3. S. Gao, C.-H. Lee and Q. Tian, "Indexing with Musical Events and Its Application to Content-Based Music Retrieval," to appear in *International Conference on Pattern Recognition* (ICPR04), Cambridge, UK, July 2004.
4. R. Shi, H. Feng, C.-H. Lee and T.-S. Chua, "An Adaptive Image Content Representation and Segmentation Approach to Automatic Image Annotation," *Proc. International Conference on Video and Image Retrieval*, Dublin, Ireland, July 2004.
5. L. Chaisorn, T.-S. Chua, C.-H. Lee and Q. Tian, "A Hierarchical Approach to Story Segmentation of Large Broadcast News Video Corpus", *Proc. International Conference on Multimedia Expo* (ICME04), Taipei, Taiwan, June 2004.
6. S. Gao, W. Wu, C.-H. Lee and T.-S. Chua, "An MFoM Learning Approach to Robust Multiclass Multi-Label Text Categorization," *International Conference on Machine Learning* (ICML04), Calgary, Alberta, July 2004.
7. S. Gao and C.-H. Lee, "An Adaptive Learning Approach to Music Tempo and Beat Analysis," *Proc. ICASSP-2004*, Montreal, Canada, April 2004.
8. S. Gao, W. Wu, C.-H. Lee and T.-S. Chua, "A Maximal Figure-of-Merit Learning Approach to Text Categorization," *2003 ACM SIGIR*, pp. 174-181, Toronto, Canada, July 2003.
9. S. Gao and C.-H. Lee, "A Hidden Markov Model Based Approach to Music Segmentation and Identification," *Proc. PCM2003*, Singapore, Dec. 2003.
10. L. Chaisorn, T.-S. Chua and C.-H. Lee, "A Multimodal Framework to Story Segmentation for News Video," *Journal of World Wide Web*, Kluwer Academic Publishers, 2003.H.-K. J. Kuo and C.-H. Lee, "Discriminative Training for Robust Natural Language Call Routing," *IEEE Trans. on Speech and Audio Proc.*, Vol. 11, No.1, pp. 24-35, Jan. 2003.
11. N. C. Maddage, C. Xu, C.-H. Lee and M. Kankanhalli, "Statistical Analysis of Musical Instruments," *Proc. PCM-2002*, Taipei, Taiwan, Dec. 2002.
12. L. Chaisorn, T.-S. Chua and C.-H. Lee, "The Segmentation of News Video into Story Units," *Proc. ICME-2002*, Lussane, Switzland, August 2002.
13. C.-H. Lee and Q. Hua, "On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition," *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1241-1269, August 2000.
14. C.-H. Lee, F. K. Soong and K. K. Paliwal (eds), *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, 1996.
15. C.-H. Lee, H, Li, L.-s. Lee, R.-H. Wang, and Q. Huo (eds), *Advances in Chinese Spoken Language Processing*, World Scientific Publishing Co., 2006.
16. R. Shi, T.-S. Chua, C.-H. Lee, and S. Gao, “'Bayesian Learning of Hierarchical Multinomial Mixture Models of Concepts for Automatic Image Annotation,” *CIVR2006*, Tempe, Arizona, July 2006.
17. J. Reed and C.-H. Lee, “A Study on Music Genre Classification Based on Universal Acoustic Models,” *Proc. ISMIR*, Victoria, BC, October 2006.
18. F. Vella and C.-H. Lee, “Information Fusuion Techniques for Automatic Image Annotation,” *Proc. VISAPP*, Barcelona, Spain, March 2007.
19. B. Byun, C.-H. Lee, S. Webb and C. Pu, “A Discriminative Classifier Learning Approach to Image Modeling and Spam Image Identification,” *Proc. CEAS*, Mountain View, CA, August 2007.
20. Y. Xiao, T.-S. Chua, L. Chaisorn, and C.-H. Lee, “Use of Generalized Pattern Model for Video Annotation,” *Proc. ICME*, Beijing, China, July 2007.

Summary

- Today's Class
  - Class project discussion
- Next Class
  - Overview on Corpus-Based Techniques
- Reading Assignments
  - M&S, Chapters 1, 2 & 3
  - *HAL's Legacy*, Chapters 6, 7 & 8