Spoken Document Recognition, Retrieval and Summarization

Berlin Chen

Associate Professor
Department of Computer Science & Information Engineering
National Taiwan Normal University

Outline

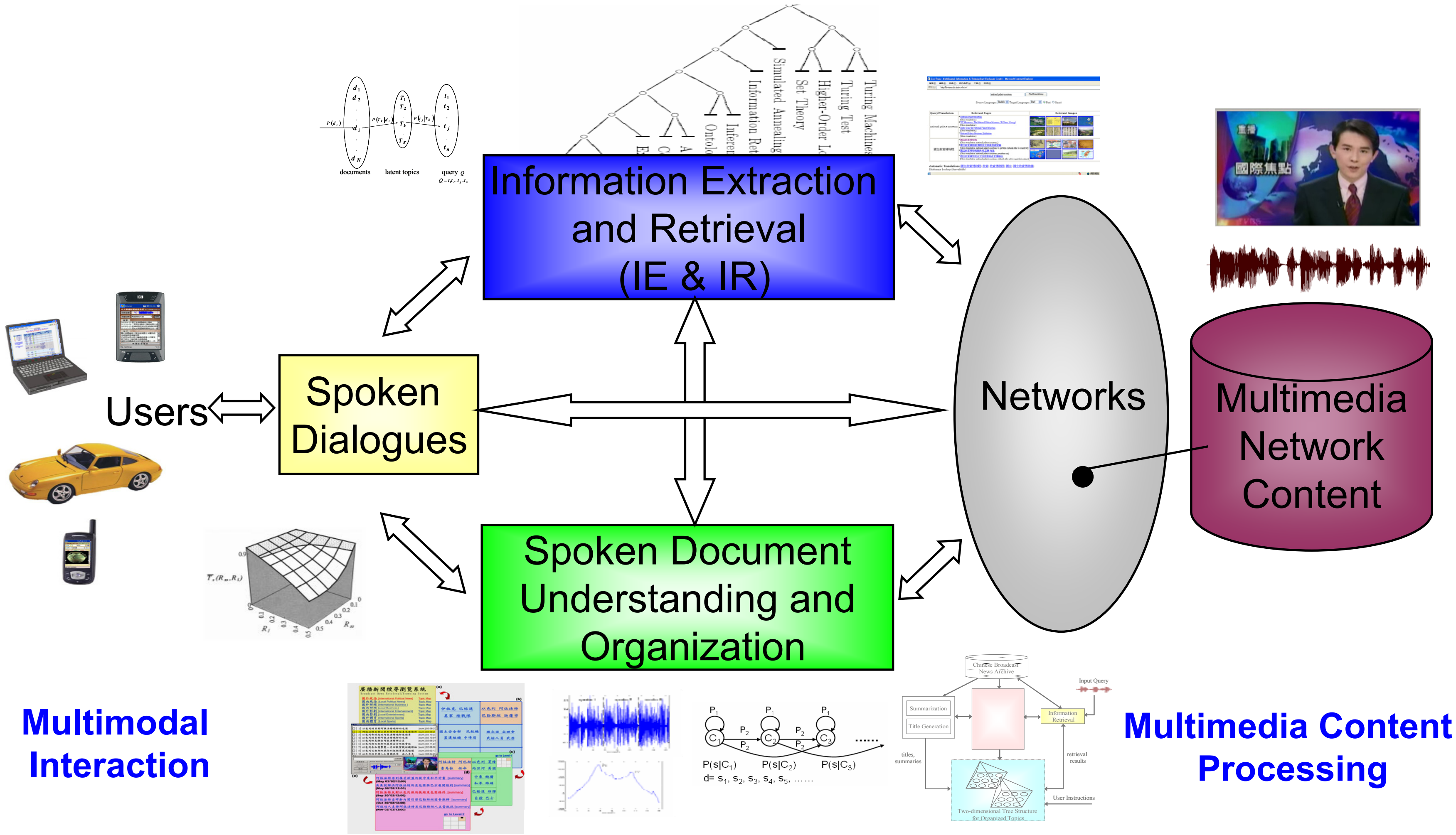
- Introduction
- Related Research Work and Applications
- Key Techniques
  - Automatic Speech Recognition (ASR)
  - Information Retrieval (IR)
  - Spoken Document Summarization
  - Spoken Document Organization
  - Named-Entity Extraction
- Prototype Systems Developed at NTNU
- Conclusions and Future Work

Introduction (1/3)

- Multimedia (audio-visual contents) associated with speech is continuously growing and filling our computers, networks and lives
  - Such as broadcast news, lectures, shows, voice mails, (contact-center) conversations, etc.
  - Speech is the most semantic (or information)-bearing
- On the other hand, speech is the primary and the most convenient means of communication between people
  - Speech provides a better (or natural) user interface in wireless environments and especially on smaller hand-held devices
- Speech will be the key for Multimedia information access in the near future

Introduction (2/3)

- Scenario for Multimedia information access using speech



Introduction (3/3)

- Organization and retrieval and of multimedia (or spoken) are much more difficult
  - Written text documents are better structured and easier to browse through
    - Provided with titles and other structure information
    - Easily shown on the screen to glance through (with visual perception)
  - Multimedia (Spoken) documents are just video (audio) signals
    - Users cannot efficiently go through each one from the beginning to the end during browsing, even if the they are automatically transcribed by automatic speech recognition
    - However, abounding speaker, emotion and scene information make them more attractive than text
    - Better approaches for efficient organization and retrieval of multimedia (spoken) documents are needed

Related Research Work and Applications

- Substantial efforts have been paid to (multimedia) spoken document recognition, organization and retrieval in the recent past [R3, R4]
  - *Informedia* System at Carnegie Mellon Univ.
  - *AT&T SCAN System*
  - *Rough'n'Ready* System at BBN Technologies
  - *SpeechBot* Audio/Video Search System at HP Labs
  - *IBM Spoken Document Retrieval for Call-Center Conversations*, *Natural Language Call-Routing, Voicemail Retrieval*
  - *NTT Speech Communication Technology for Contact Centers*
  - *Google Voice Local Search*

Key Techniques (1/2) [R4]

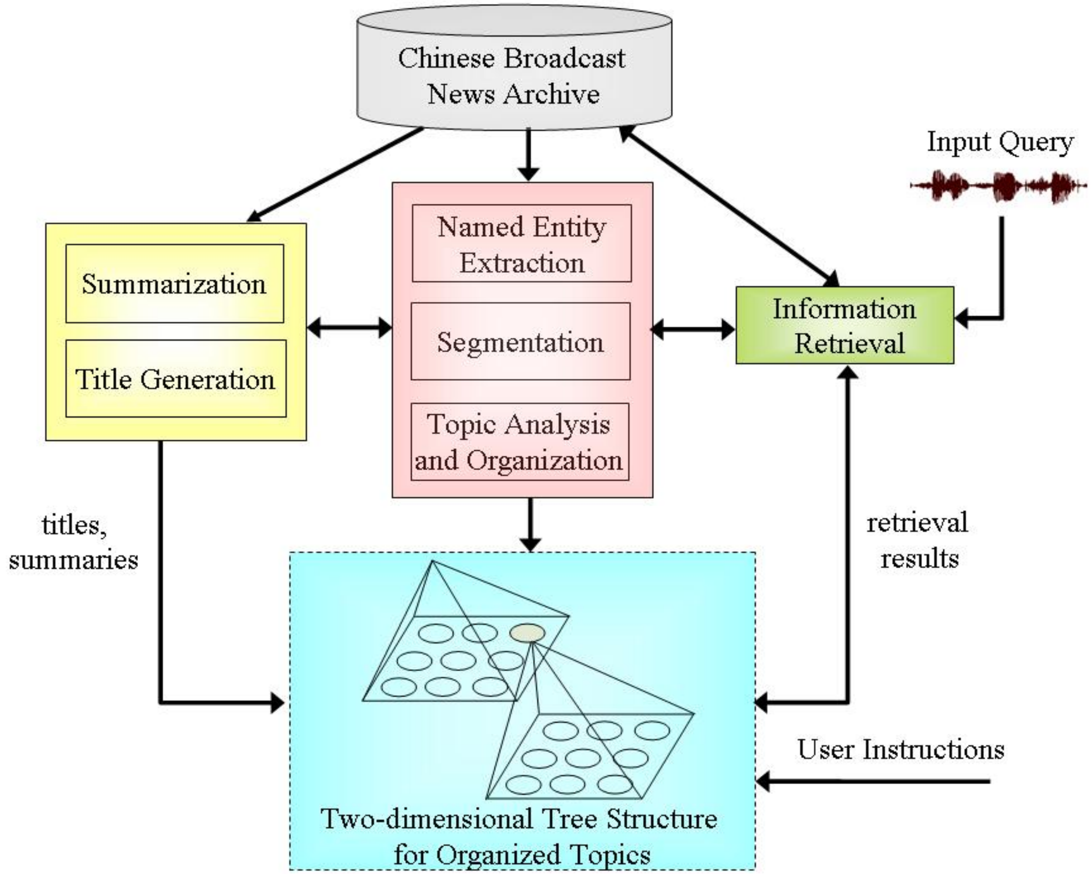
- Automatic Speech Recognition
  - Automatically convert speech signals into sequences of words or other suitable units for further processing
- Spoken Document Segmentation
  - Automatically segment speech signals (or automatically transcribed word sequences) into a set of documents (or short paragraphs) each of which has a central topic
- Audio Indexing and Information Retrieval
  - Robust representation of the spoken documents
  - Matching between (spoken) queries and spoken documents
- Named Entity Extraction from Spoken Documents
  - Personal names, organization names, location names, event names
  - Very often out-of-vocabulary (OOV) words, difficult for recognition

Key Techniques (2/2)

- Information Extraction for Spoken Documents
  - Extraction of key information such as who, when, where, what and how for the information described by spoken documents
- Summarization for Spoken Documents
  - Automatically generate a summary (in text or speech form) for each spoken document or a set of topic-coherent documents
- Title Generation for Multi-media/Spoken Documents
  - Automatically generate a title (in text/speech form) for each short document; i.e., a very concise summary indicating the themes of the documents
- Topic Analysis and Organization for Spoken Documents
  - Analyze the subject topics for (retrieved) documents
  - Organize the subject topics of documents into graphic structures for efficient browsing

An Example System for Chinese Broadcast News (1/2)

- For example, a prototype system developed at NTU for efficient spoken document retrieval and browsing [R4]



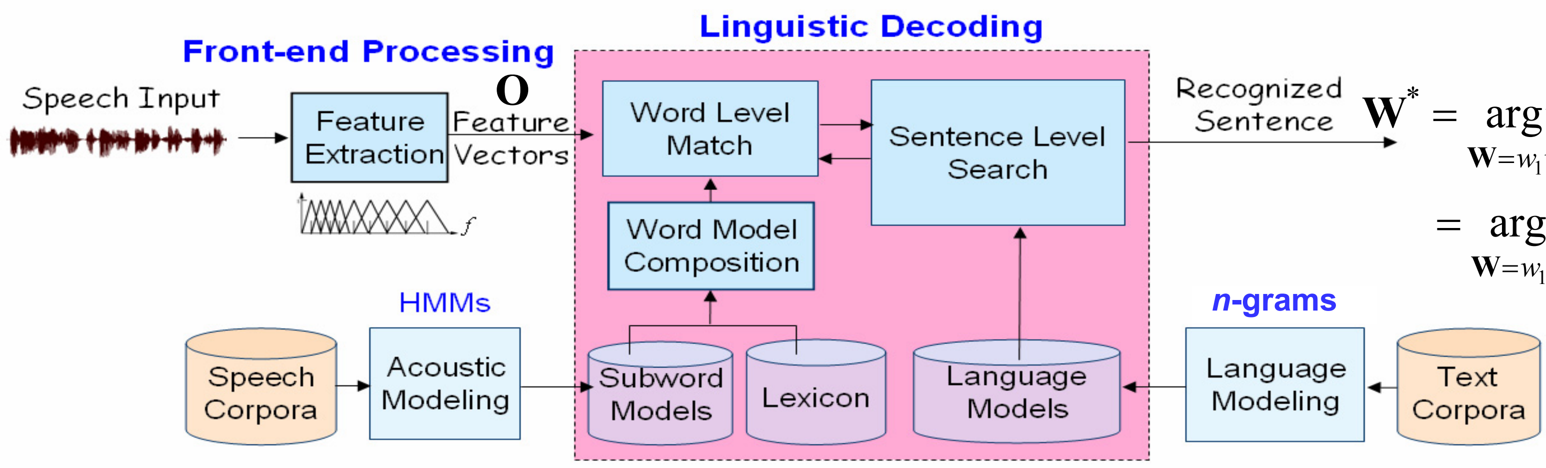
An Example System for Chinese Broadcast News (2/2)

- Users can browse spoken documents in top-down and bottom-up manners

http://sovideo.iis.sinica.edu.tw/NeGSST/Index.htm

Automatic Speech Recognition (1/3)

- Large Vocabulary Continuous Speech Recognition (LVCSR)



$$\mathbf{W}^* = \underset{\mathbf{W}=w_1 w_2 \ldots w_L}{\operatorname{argmax}} P(\mathbf{W}|\mathbf{O})$$
$$= \underset{\mathbf{W}=w_1 w_2 \ldots w_L}{\operatorname{argmax}} P(\mathbf{O}|\mathbf{W})P(\mathbf{W})$$

- The speech signal is converted into a sequence of feature vectors
- The pronunciation lexicon is structured as a tree
- Due to the constraints of *n*-gram language modeling, a word's occurrence is dependent on its previous *n*-1 words

$$P(w_i|w_1 w_2 \ldots w_{i-1}) \approx P(w_i|w_{i-n+1} w_{i-n+2} \ldots w_{i-1})$$

- Search through all possible lexical tree copies from the start time to the end time of the utterance to find the best sequence among the word hypotheses

Automatic Speech Recognition (2/3)

- Discriminative and Robust Speech Feature Extraction
  - Heteroscedastic Linear Discriminant Analysis (HLDA) and Maximum Likelihood Linear Transformation (MLLT) for distriminative speech feature extraction
  - Polynomial-fit Histogram Equalization (PHEQ)aApproaches for obust speech feature extraction *Interspeech 2006, 2007; ICME 2007; ASRU 2007*
- Acoustic Modeling
  - Lightly-Supervised Training of Acoustic Models *ICASSP 2004*
  - Data Selection for Discriminative Training of Acoustic Models (HMMs) *ICME 2007; ASRU 2007*
- Dynamic Language Model Adaptation
  - Minimum Word Error (MWE) Training *Interspeech 2005*
  - Word Topical Mixture Models (WTMM) *ICASSP 2007*
- Linguistic Decoding
  - Syllable-level acoustic model look-ahead *ICASSP 2004*

Automatic Speech Recognition (3/3)

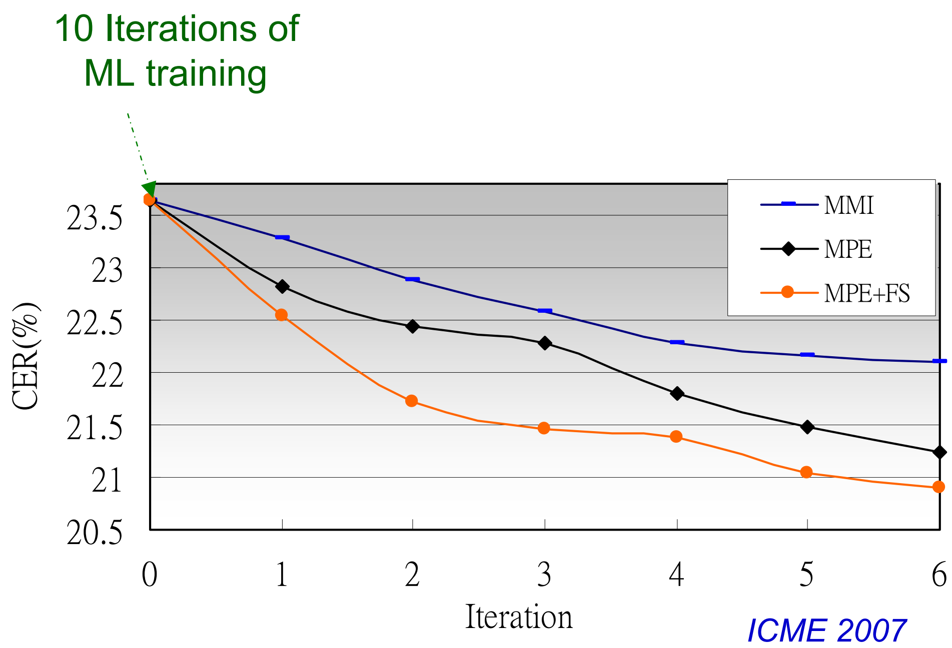
- Transcription of PTS (Taiwan) Broadcast News

Automatic

根據 最新 但 雨量 統計
一 整天 下來
費 雪 以 試辦 兩 個 水庫 的 雨量
分別 是 五十 三 公里 和 二十 九 公厘
對 水位 上升 幫助 不 大
不過 就業機會 期間 也 多 在 夜間
氣象局 也 針對 中部以北 及 東北部 地區 發佈 豪雨 特 報
因此 還是 有 機會 增加 積水 區 的 降雨量
此外 氣象局 也 預測
華航 又 有 另 一道 鋒面 通過
水利署 估計 如果 這 波 鋒面 能 帶來 跟著 會 差不多 的 雨水
那 個 北 台灣 的 第二 階段 限水 時間
渴望 見到 五月 以後
公視 新聞 當時 匯率 採訪 報導

Manual

根據最新的雨量統計
一整天下來
翡翠石門兩個水庫的雨量
分別是五十三公厘和二十九公厘
對水位上升幫助不大
不過由於集水區降雨多在夜間
氣象局也針對中部以北及東北部地區發布了豪雨特報
因此還是有機會增加集水區的降雨量
此外氣象局也預測
八號又有另一道鋒面通過
水利署估計如果這波鋒面能帶來跟這回差不多的雨水
那麼北台灣的第二階段限水時間
可望延到五月以後
公視新聞張玉菁陳柏諭採訪報導



Relative Character Error Rate Reduction
- MMI: 6.5%
- MPE: 10.1%
- MPE+FS: 11.6%

Information Retrieval Models

- Information retrieval (IR) models can be characterized by two different matching strategies
  - Literal term matching
    - Match queries and documents in an index term space
  - Concept matching
    - Match queries and documents in a latent semantic space

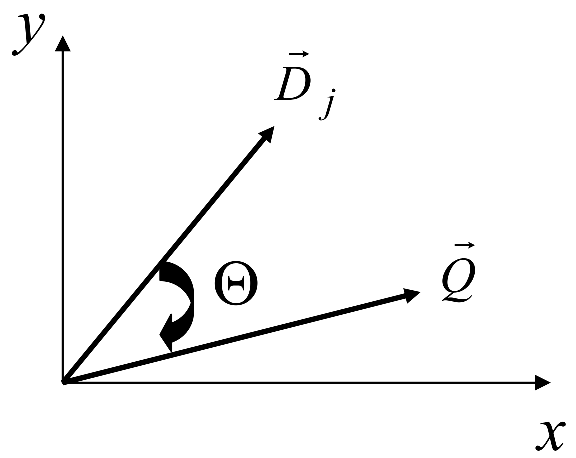
relevant ?

香港星島日報篇報導引述軍事觀察家的話表示，到二零零五年台灣將完全喪失空中優勢，原因是中國大陸戰機不論是數量或是性能上都將超越台灣，報導指出中國在大量引進俄羅斯先進武器的同時也得加快研發自製武器系統，目前西安飛機製造廠任職的改進型飛豹戰機即將部署尚未與蘇愷三十通道地對地攻擊住宅飛機，以督促遇到挫折的監控其戰機目前也已經取得了重大階段性的認知成果。根據日本媒體報導在台海戰爭隨時可能爆發情況之下北京方面的基本方針，使用高科技答應局部戰爭。因此，解放軍打算在二零零四年前又有包括蘇愷三十二期在內的兩百架蘇霍伊戰鬥機。

IR Models: Literal Term Matching (1/2)

- Vector Space Model (VSM)
  - Vector representations are used for queries and documents
  - Each dimension is associated with a index term (TF-IDF weighting)
  - Cosine measure for query-document relevance

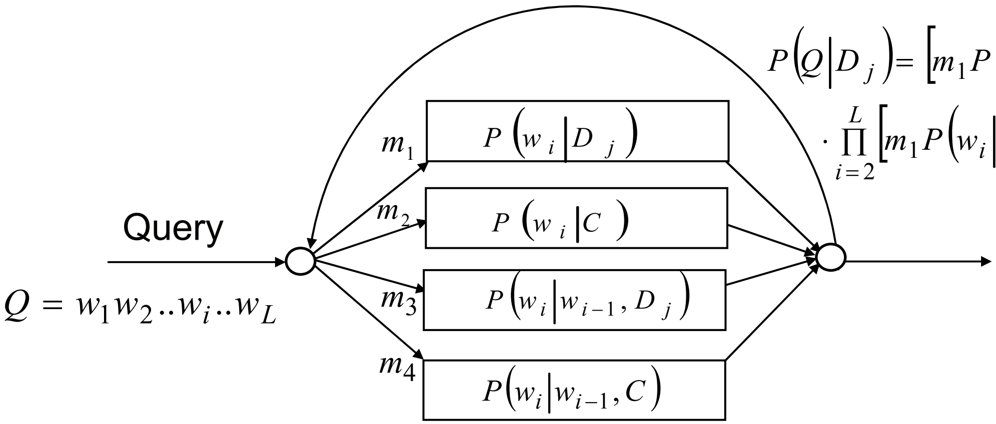
$$sim(D_j, Q) = \text{cosine}(\Theta) = \frac{\vec{D}_j \bullet \vec{Q}}{|\vec{D}_j| \times |\vec{Q}|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,Q}^2}}$$



  - VSM can be implemented with an inverted file structure for efficient document search (instead of exhaustive search)

IR Models: Literal Term Matching (2/2)

- Hidden Markov Model (HMM) [R1]
  - Also thought of as Language Model (LM)
  - Each document is a probabilistic generative model consisting of a set of *N*-gram distributions for predicting the query

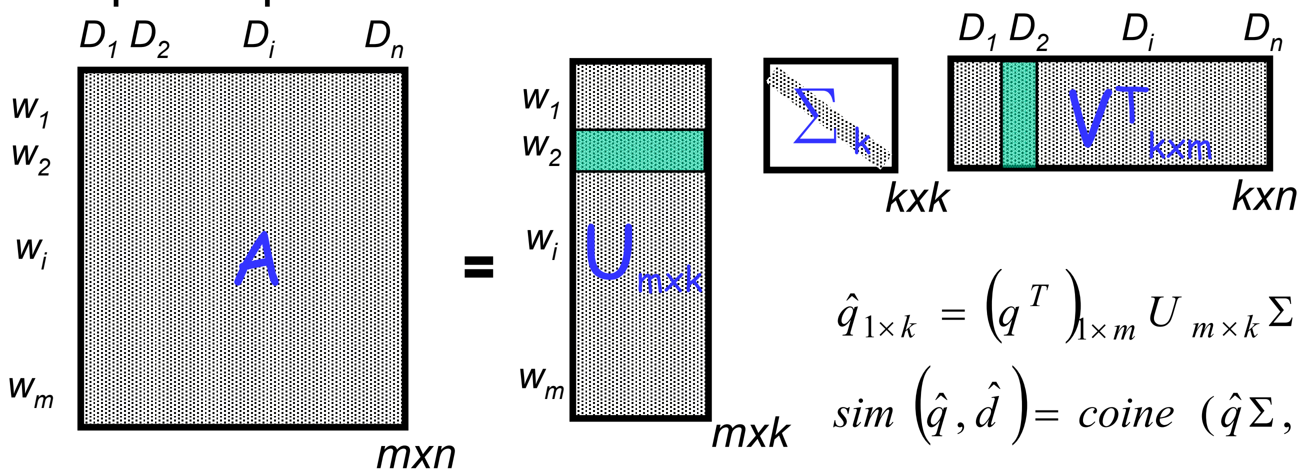


$$P(Q|D_j) = [m_1 P(w_1|D_j) + m_2 P(w_1|C)] \cdot \prod_{i=2}^{L} [m_1 P(w_i|D_j) + m_2 P(w_i|C) + m_3 P(w_i|w_{i-1}, D_j) + m_4 P(w_i|w_{i-1}, C)]$$

  - Models can be optimized by the expectation-maximization (EM) or minimum classification error (MCE) training algorithms
  - Such approaches do provide a potentially effective and theoretically attractive probabilistic framework for studying IR problems

IR Models: Concept Matching (1/3)

- Latent Semantic Analysis (LSA) [R2]
  - Start with a matrix describing the intra- and Inter-document statistics between all terms and all documents
  - Singular value decomposition (SVD) is then performed on the matrix to project all term and document vectors onto a reduced latent topical space
  - Matching between queries and documents can be carried out in this topical space

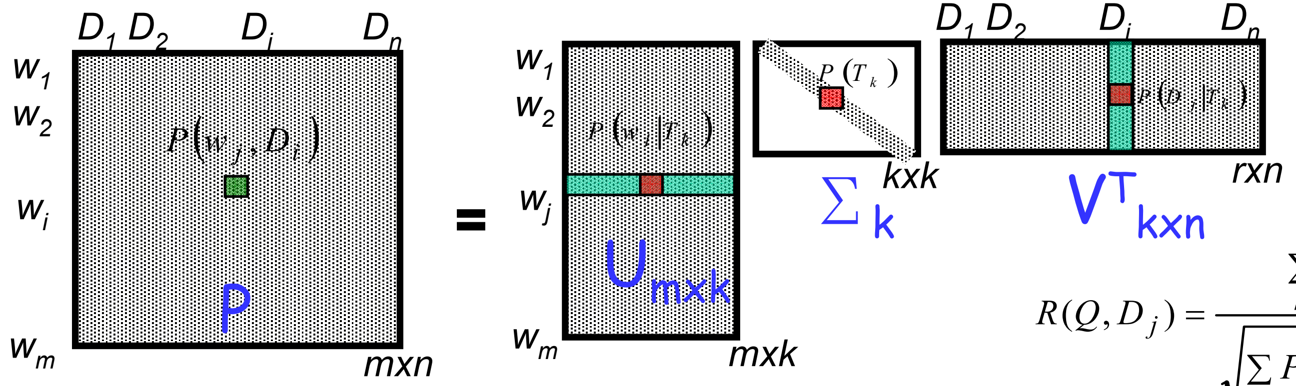


$$\hat{q}_{1\times k} = \left(q^{T}\right)_{1\times m} U_{m\times k} \Sigma^{-1}_{k\times k}$$

$$sim\left(\hat{q},\hat{d}\right) = coine\left(\hat{q}\Sigma,\hat{d}\Sigma\right) = \frac{\hat{q}\Sigma^{2}\hat{d}^{T}}{\left|\hat{q}\Sigma\right|\left|\hat{d}\Sigma\right|}$$

IR Models: Concept Matching (2/3)

- Probabilistic Latent Semantic Analysis (PLSA) [R5, R6]
  - An probabilistic framework for the above topical approach



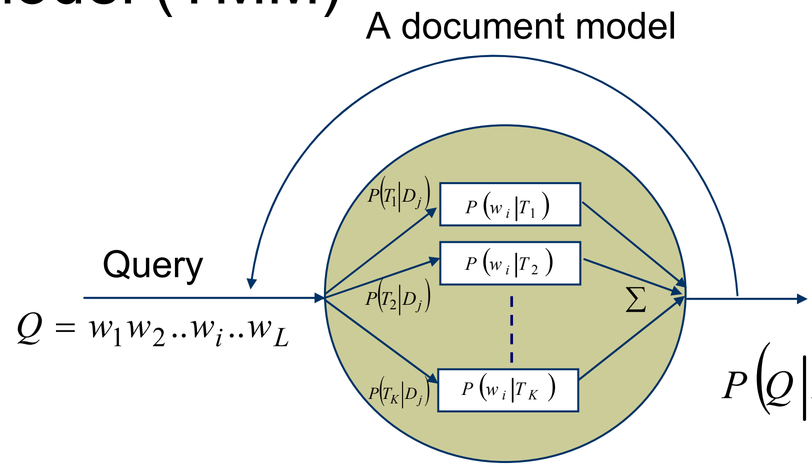
$$R(Q, D_j) = \frac{\sum_k P(T_k|Q)P(T_k|D_j)}{\sqrt{\sum_k P(T_k|Q)^2}\sqrt{\sum_k P(T_k|D_j)^2}}$$

$$P(w_i, D_j) \approx \sum_k P(w_i|T_k)P(T_k)P(D_j|T_k)$$

  - Relevance measure is not obtained directly from the frequency of a respective query term occurring in a document, but has to do with the frequency of the term and document in the latent topics
  - A query and a document thus may have a high relevance score even if they do not share any terms in common

IR Models: Concept Matching (3/3)

- PLSA also can be viewed as an HMM model or a topical mixture model (TMM)



$$P(Q|D_j) = \prod_{i=1}^{L} \left[\sum_{k=1}^{K} P(w_i|T_k) P(T_k|D_j) \right]$$

  - Explicitly interpret the document as a mixture model used to predict the query, which can be easily related to the conventional HMM modeling approaches widely studied in speech processing community (topical distributions are tied among documents)
  - Thus quite a few of theoretically attractive model training algorithms can be applied in supervised or unsupervised manners

IR Evaluations

- Experiments were conducted on TDT2/TDT3 spoken document collections [R6]
  - TDT2 for parameter tuning/training, while TDT3 for evaluation
  - E.g., mean average precision (*m*AP) tested on TDT3

	VSM	LSA	TMM	HMM	PLSA
TD	0.6505	0.6440	0.7870	0.7174	0.6882
SD	0.6216	0.6390	0.7852	0.7156	0.6688

TALIP2004; Interspeech2004, 2005; PATREC 2006

    - HMM/PLSA/TMM are trained in a supervised manner

- Language modeling approaches (TMM/PLSA/HMM) are evidenced with significantly better results than that of conventional statistical approaches (VSM/LSA) in the above spoken document retrieval (SDR) task

Spoken Document Summarization (1/2)

- Spoken document summarization (SDS), aiming to generate a summary automatically for the spoken documents, is the key for better speech understanding and organization

- Extractive vs. Abstractive Summarization
  - **Extractive summarization** is to select a number of indicative sentences or paragraphs from original document and sequence them to form a summary
  - **Abstractive summarization** is to rewrite a concise abstract that can reflect the key concepts of the document
  - Extractive summarization has gained much more attention in the recent past

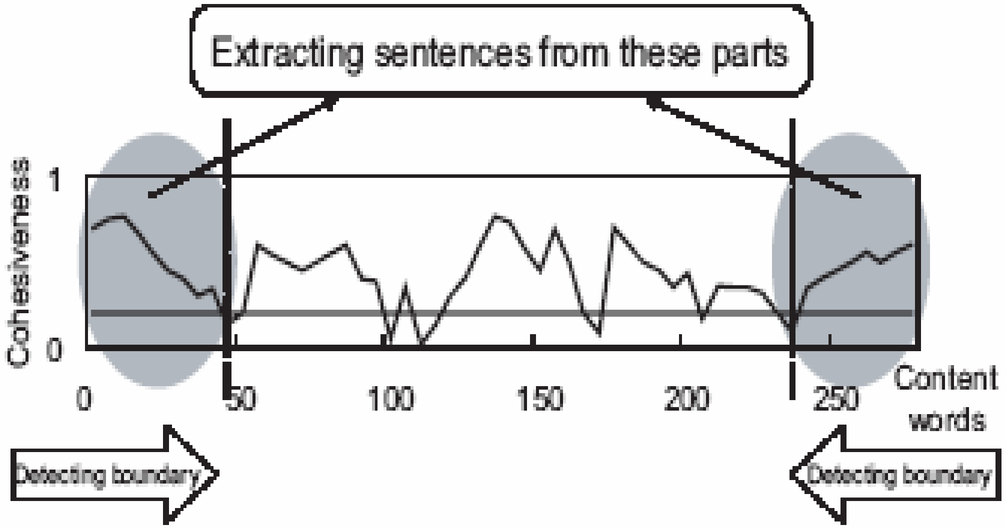
Spoken Document Summarization (2/2)

- Common Extractive Document Summarization Approaches
  - Based on based on sentence structure or location information
  - Based on statistical measures
  - Based on sentence classification
  - Based on sentence generative probabilities

  - There has also been some research on exploring
    - Extra information clues, e.g.
      - word-clusters, WordNet, or event relevance
    - Novel ranking algorithms

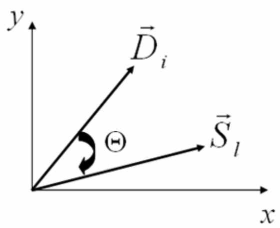
SDS: Approaches Based on Sentence Structure or Location Information

- Lead (Hajime and Manabu 2000)
- Focus on the introductory and concluding segments (Hirohata et al. 2005)
- Specific structure on some domain (Maskey et al. 2003)
  – E.g., broadcast news programs－sentence position, speaker type, previous-speaker type, next-speaker type, speaker change

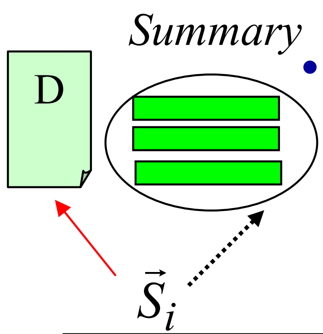


SDS: Approaches Based on Statistical Measures (1/3)

- Vector Space Model (VSM) Y. Gong, *SIGIR 2001*
  - Vector representations of sentences and the document to be summarized using statistical weighting such as *TF-IDF*
  - Sentences are ranked based on their proximity to the document
  - To summarize more important and different concepts in a document



    - The terms occurring in the sentence with the highest relevance score $Sim(S_l, D_i)$ are removed from the document
    - The document vector is then reconstructed and the ranking of the rest of the sentences is performed accordingly



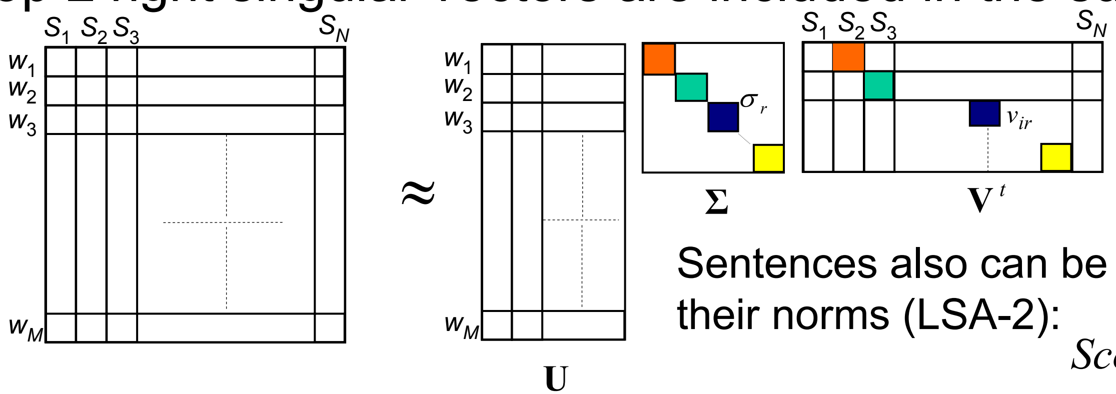
- Or, using the Maximum Marginal Relevance (MMR) model

$$NextSen = \max_{S_l} \left[\lambda \cdot Sim(S_l, D_i) - (1-\lambda) Sim(S_l, Summ) \right]$$

  - $Summ$: the set of already selected sentences

SDS: Approaches Based on Statistical Measures (2/3)

- Latent Semantic Analysis (LSA) Y. Gong, *SIGIR 2001*
  - Construct a “term-sentence” matrix for a given document
  - Perform SVD on the “term-sentence” matrix
    - The **right singular vectors** with larger singular values represent the dimensions of the more important latent semantic concepts in the document
    - Represent each sentence of a document as a vector in the latent semantic space
  - Sentences with the largest index (element) values in each of the top *L* right singular vectors are included in the summary (LSA-1)



Hirohata et al., ICASSP2005

Sentences also can be selected based on their norms (LSA-2):

$$Score(S_i) = \sqrt{\sum_{r=1}^{L} (\sigma_r v_{ir})^2}$$

SDS: Approaches Based on Statistical Measures (3/3)

- Sentence Significance Score (SIG)
  - Sentences are ranked based on their significance which, for example, is defined by the average importance scores of words in the sentence

$$SIG(S_i) = \frac{1}{N_S} \sum_{n=1}^{N_S} I(w_n)$$

$$I(w_n) = f_w \cdot icf = f_w \cdot \log \frac{F_C}{F_w}$$

similar to *TF-IDF* weighting

S. Furui et al., IEEE SAP 12(4), 2004

  - Other features such as *word confidence*, *linguistic score*, or *prosodic information* also can be further integrated into this method

$$SIG(S_i) = \frac{1}{N_{S_i}} \sum_{n=1}^{N_{S_i}} \{\lambda_1 s(w_n) + \lambda_2 l(w_n) + \lambda_3 c(w_n) + \lambda_4 g(w_n)\} + \lambda_5 b(S_i)$$

$s(w_n)$:statistical measure, such as TF/IDF

$l(w_n)$:linguistic measure, e.g., named entities and POSs

$c(w_n)$:confidence score

$g(w_n)$:N-gram score

$b(S_i)$:is calculated from the grammatical structure of the sentence

SDS: Approaches Based on Sentence Classification (1/2)

- Sentence selection is formulated as a binary classification problem
  - A sentence can either be included in a summary or not
- A bulk of classification-based methods using statistical features also have been developed
  - Gaussian Mixture Models (GMM)
  - Bayesian Network (BN)
  - Support Vector Machine (SVM)
  - Logistic Regression (LR)
  - Conditional Random Fields (CRFs)

$$P(S_i \in \mathbf{S} \mid X_i) = \frac{p(X_i \mid S_i \in \mathbf{S}) P(S_i \in \mathbf{S})}{P(X_i)}.$$

Summary

Non-summary

S_i

- However, the above methods need a set of training documents together with their corresponding handcrafted summaries (or labeled data) for training the classifiers

SDS: Approaches Based on Sentence Classification (2/2)

- Example of a set of features used in the classification-based methods

Structural features	*POSITION*: Sentence position *DURATION*: Duration of preceding/current/following sentence
Lexical Features	*BIGRAM_SCORE*: Normalized bigram language model scores *SIMILARITY*: Similarity scores between a sentence and its preceding/following neighbors *NUM_NAME_ENTITES*: Number of name entities (NE) in a sentence
Acoustic Features	*PITCH*: Min/max/mean/difference pitch values of a spoken sentence *ENERGY*: Min/max/mean/difference value of energy features of a spoken sentence *CONFIDENCE*: Posterior probabilities
Relevance Features	*VSM*: Relevance score obtain by using the VSM summarizer *LSA*: Relevance score obtain by using the LSA summarizer

SDS: Approaches Based on Sentence Generative Probabilities (1/2)

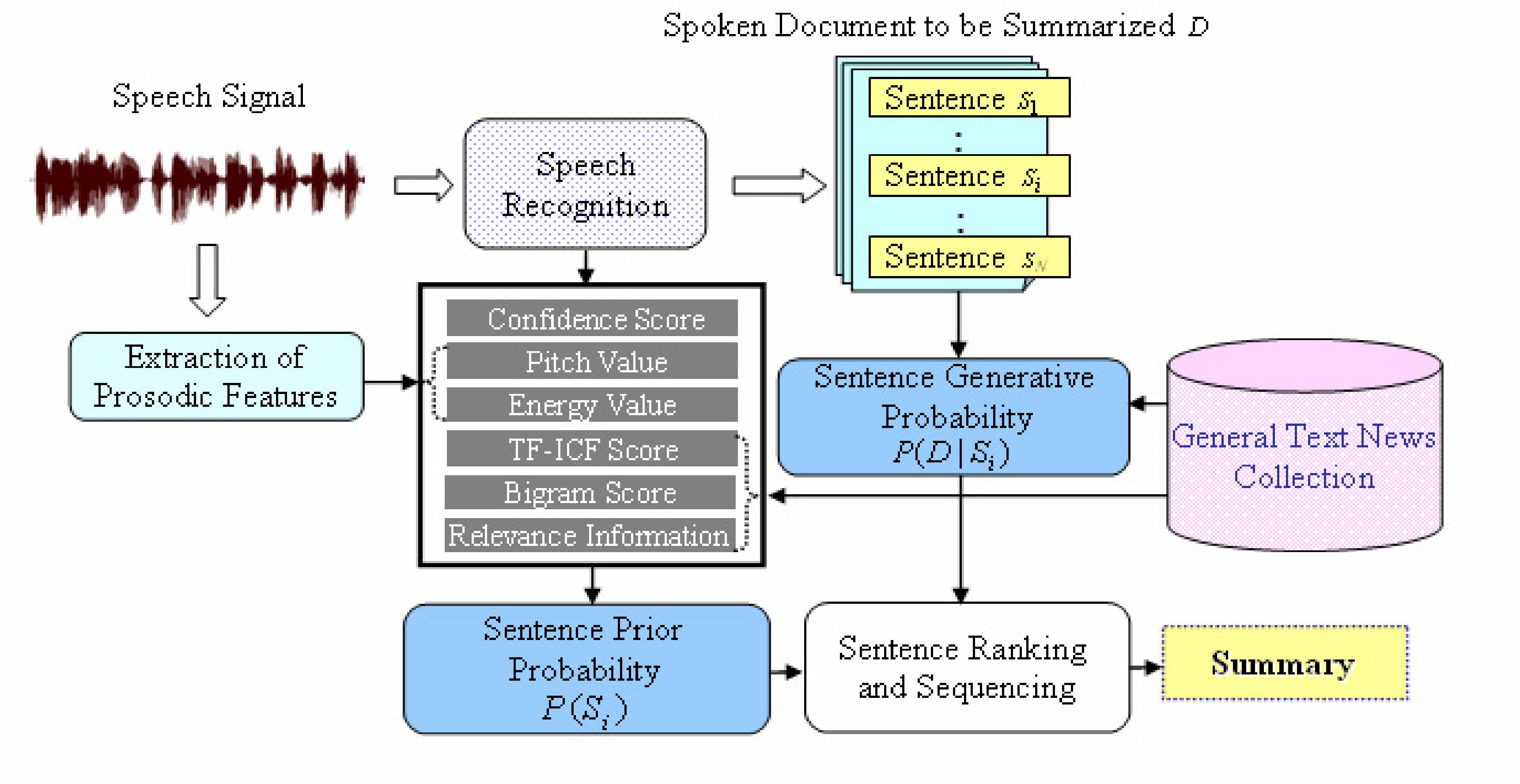
- A **Probabilistic Generative Framework** for Sentence Selection (Ranking)
  - Maximum a Posteriori Probability (MAP) Criterion

$$P(S_i|D) = \frac{P(D|S_i)P(S_i)}{P(D)} \propto P(D|S_i)P(S_i)$$

- Sentence Generative Model, $P(D|S_i)$
  - Each sentence of the document as a probabilistic generative model
  - Language Model (LM), Sentence Topical Mixture Model (STMM) and Word Topical Mixture Model (WTMM) are initially investigated
    ICASSP2006; ISCSLP2006; ICME 2007; PATREC 2007; ICASSP2008
- Sentence Prior Distribution, $P(S_i)$
  - The sentence prior distribution may have to do with sentence *duration/position*, *correctness of sentence boundary*, *confidence score*, *prosodic information*, etc. (e.g., they can be fused by the whole-sentence maximum entropy model) *Interspeech2007; ASRU 2007*

SDS: Approaches Based on Sentence Generative Probabilities (2/2)

- A flowchart for our proposed framework



SDS: Evaluation Metrics (1/2)

- Subjective Evaluation Metrics (direct evaluation)
  - Conducted by human subjects
  - Different levels
- Objective Evaluation Metrics
  - Automatic summaries were evaluated by objective metrics
- Automatic Evaluation
  - Summaries are evaluated by IR

SDS: Evaluation Metrics (2/2)

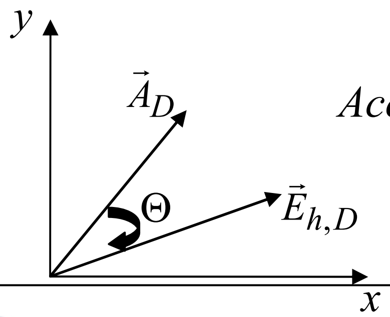
- Objective Evaluation Metrics
  - **ROUGE-*N*** (Lin et al. 2003)
    - ROUGE-*N* is an *N*-gram recall between an automatic summary and a set of manual summaries

$$\text{ROUGE}-N = \frac{\sum_{\mathbf{S}\in\mathbf{S}_H}\sum_{g_N\in\mathbf{S}} C_m(g_N)}{\sum_{\mathbf{S}\in\mathbf{S}_H}\sum_{g_N\in\mathbf{S}} C(g_N)}$$

S_H : a set of human summaries

$C_m(g_N)$: number of matched N-grams between human and automatic summary

  - **Cosine Measure** (Saggion et al. 2002)



$$Acc_D = \frac{1}{2}\left[sim(E, E_R) + sim(E, A_R)\right]$$

E : automatic extractive summary

E_R : reference extractive summary

A_R : reference abstractive summary

SDS: Experimental Results

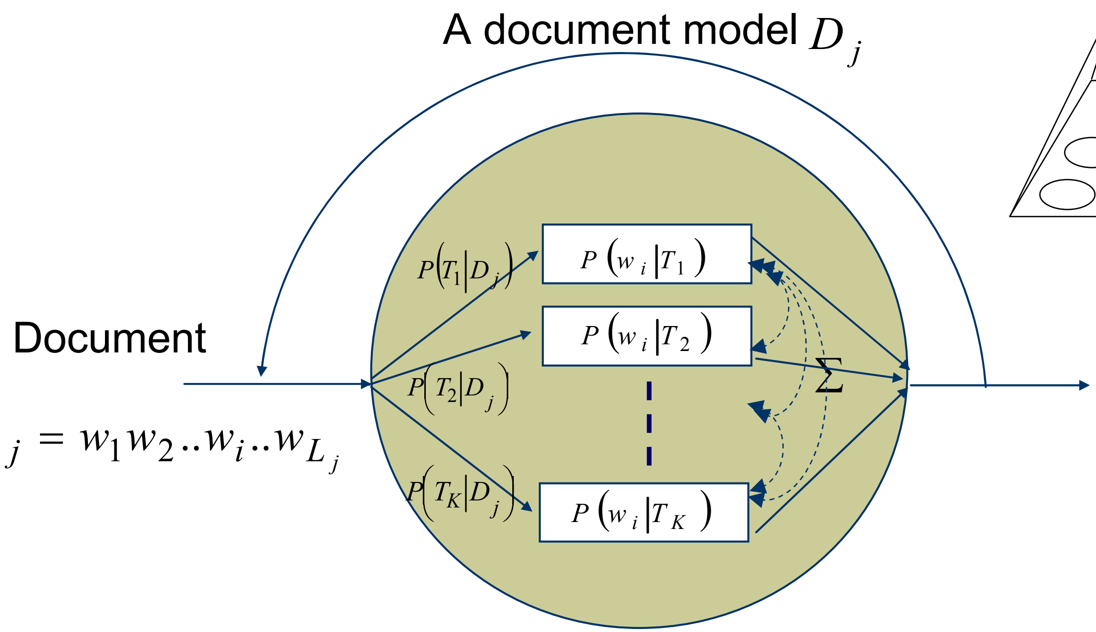
- Preliminary tests on 100 radio broadcast news stories collected in Taiwan (automatic transcripts with 14.17% character error rate)
  - ROUGE-2 measure was used to evaluate the performance levels of different models

	VSM	MMR	LSA	DIM	SIG	SVM	LM-RT	WTMM
10%	0.3073	0.3073	0.3034	0.3187	0.3144	0.3425	0.3684	0.3836
20%	0.3188	0.3214	0.2926	0.3148	0.3259	0.3408	0.3696	0.3772
30%	0.3593	0.3678	0.3286	0.3383	0.3428	0.3719	0.3840	0.3728
50%	0.4485	0.4501	0.3906	0.4345	0.4666	0.4660	0.4884	0.4615

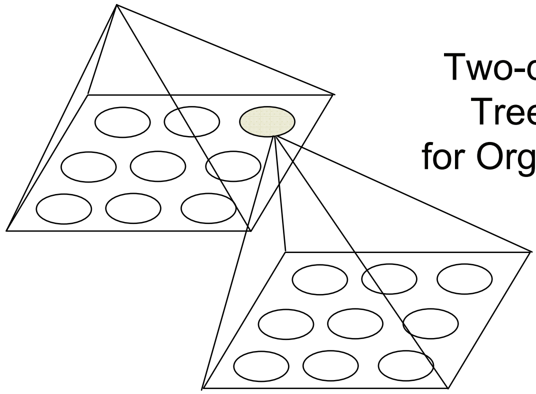
  - Proposed models (LM, WTMM) are consistently better than the other models at lower summarization ratios
    - LM and WTMM are trained in a pure unsupervised manner, without any document-summary relevance information (labeled data)

Spoken Document Organization (1/3)

- Each document is viewed as a TMM model to generate itself
  - Additional transitions between topical mixtures have to do with the topological relationships between topical classes on a 2-D map



$$D_j = w_1 w_2 .. w_i .. w_{L_j}$$



$$E(T_l, T_k) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{dist(T_k, T_l)^2}{2\sigma^2} \right]$$

$$P(T_l|T_k) = \frac{E(T_l, T_k)}{\sum_{s=1}^{K} E(T_s, T_k)}$$

$$P(w_i|D_i) = \sum_{k=1}^{K} P(T_k|D_i) \left[\sum_{l=1}^{K} \underline{P(T_l|T_k)} P(w_n|T_l) \right]$$

Spoken Document Organization (2/3)

- Document models can be trained in an unsupervised way by maximizing the total log-likelihood of the document collection

$$L_T = \sum_{j=1}^{n} \sum_{i=1}^{V} c\left(w_i, D_j\right) \log P\left(w_i \middle| D_j\right)$$

- Initial evaluation results (conducted on the TDT2 collection)

Model	Iterations	$dist_{Bet}/dist_{Within}$
TMM	10	1.9165
SOM	100	2.0604

  – TMM-based approach is competitive to the conventional Self-Organization Map (SOM) approach

Spoken Document Organization (3/3)

- Each topical class can be labeled by words selected using the following criterion

$$Sig\left(w_i, T_k\right) = \frac{\sum_{j=1}^{n} c\left(w_i, D_j\right) P\left(T_k \middle| D_j\right)}{\sum_{i=1}^{n} c\left(w_i, D_j\right)\left[1 - P\left(T_k \middle| D_j\right)\right]}$$

- An example map for international political news

聯邦調查局 執法 劃歸 空對空飛彈 安全部 艾希克羅 蓋達組織 接種 等級 民航機 認出 輻射性 劫機 主謀 重整旗鼓 歐瑪 穆勒 國土 黃色 葉門 美國境內 中情局 天花 芮吉	僑界 僑務 台商 會長 僑胞 呼吸 雙十國慶 酒會 立委 舉辦 國慶 聯誼會 經文 履新 組長 衛生 餐會 春節 渥太華 後援 中華 僑團 華僑 鄉親	法輪 鈴木宗男 巫統 中國共產黨 李光耀 挪用 書記 交替 班子 馬哈地 一邊 李顯龍 吳作棟 新疆 論說 軍委 政治局 標題 馬來人 早報 格局 資政 接班 報章
檢查人員 檢查員 動武 最後通牒 安理會 布里克斯 決議 精密 武檢 聯合國 授權 沙丹． 銷毀 違禁 解除 武檢人員 檢查 首席 武器 決議案 胡笙 禁航區 導引 毀滅性	西非 衛隊 巴格達機場 伊拉克部隊 伊拉克南部 賴比瑞亞 伊北 科威特 步兵 辛格 庫德族 斯拉 法新社 翁山蘇姬 庫克 蒙羅維亞 巴格達 陸戰隊 轟炸 激戰 卡達 克里 市中心 基爾	林東源 金大中 漢城 南北 多邊 正常化 長官 平壤 分界線 會談 鐵路 南韓統一部 韓美 燃料 南韓 懸案 金正日 盧武鉉 朝鮮半島 打撈 黃海 銜接 核子 北韓
普查 支領 王太 王室 登基 會計年度 小泉內閣 瑪格麗特 問卷 靈柩 溫莎堡 英鎊 西敏寺大廳 白金漢宮 社會勞工黨 王太后 加班 女王 降至 百分點 享年 伊麗莎白 太后 大關	自殺 加薩市 炸彈 巴勒斯坦 城鎮 約旦河 巴勒斯坦人 哈瑪斯 喪生 耶路撒冷 阿拉法特 約旦河西岸 以色列 伯利恆 槍手 加薩走廊 夏隆 墾區 西岸 受傷 特拉維夫 以色列部隊 包圍 巴士	中美洲 決選 薩爾瓦多 哥斯大黎加 中間 兼職 雷朋 宏都拉斯 羅育 馬達加斯加 史瓦濟蘭 翁岳生 王金平 勳章 院長 金哥納 馬拉坎南宮 游錫方 右派 雅羅 查維斯 喬斯班 孟代爾 方土

Named-Entity Extraction (1/10)

- Named entities (NE) include
  - Proper nouns as names for persons, locations, organizations, artifacts and so on
  - Temporal expressions such as “Oct. 10 2003” or “1:40 p.m.”
  - Numerical quantities such as “fifty dollars” or “thirty percent”

- Temporal expressions and numerical quantities can be easily modeled and extracted by rules

- The personal/location/organization are much more difficult to identified
  - E.g., “White House” can be either an organization or a location name in different context

Named-Entity Extraction (2/10)

- NE has it origin from the Message Understanding Conferences (MUC) sponsored by U.S. DARPA program
  - Began in the 1990's
  - Aimed at extraction of information from text documents
  - Extended to many other languages and spoken documents (mainly broadcast news)

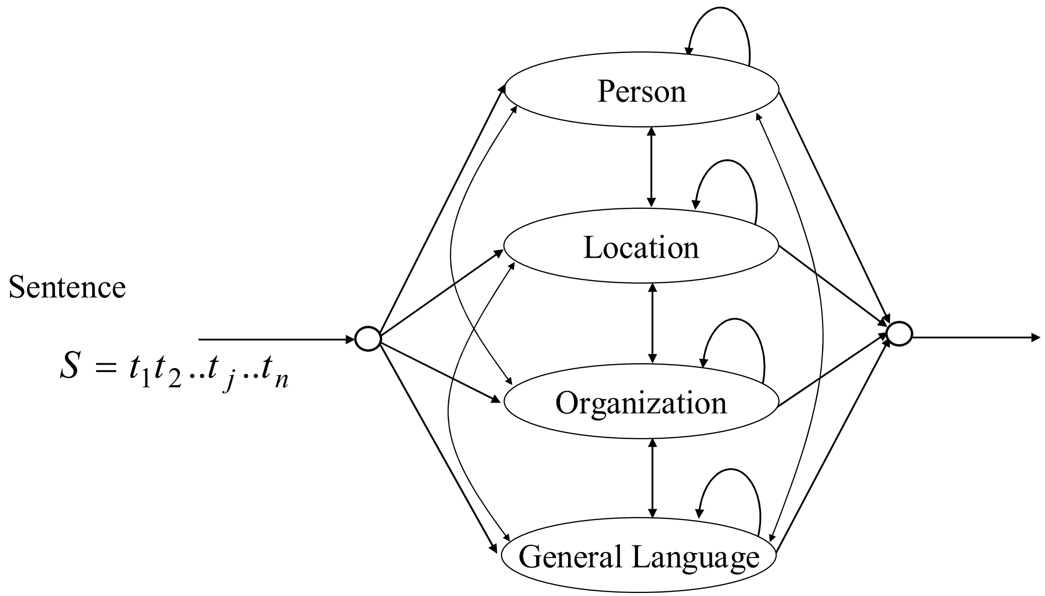
- Common approaches to NE
  - Rule-based approach
  - Model-based approach
  - Combined approach

Named-Entity Extraction (3/10)

- Rule-based Approach
  - Employ various kinds of rules to identified named-entities
  - E.g.,
    - A cue-word “Co.” possibly indicates the existence of a company name in the span of its predecessor words
    - A cue-word “Mr.” possibly indicates the existence of a personal name in the span of its successor words
  - However, the rules may become very complicated when we wish to cover all different possibilities
    - Time-consuming and difficult to handcraft all the rules
    - Especially when the task domain becomes more general, or when new sources of documents are being handled

Named-Entity Extraction (4/10)

- Model-based Approach
  - The goal is usually to find the sequence of named entity labels (personal name, location name, etc.), $E = e_1 e_2 ..e_j ..e_n$, for a sentence, $S = t_1 t_2 ..t_j ..t_n$, which maximizes the probability $P(E|S)$
  - E.g., HMM is probably the best typical representative model used in this category



Named-Entity Extraction (5/10)

- In HMM,
  - One state modeling each type of the named entities (person, location, organization)
  - One state modeling other words in the general language (non-named-entity words)
  - Possible transitions from states to states
  - Each state is characterized by a bi- or trigram language model
  - Viterbi search to find the most likely state sequence, or named entity label sequence E , for the input sentence, and the segment of consecutive words in the same named entity state is taken as a named entity

Named-Entity Extraction (6/10)

- Combined approach
  - E.g., Maximum entropy (ME) method
    - Many different linguistic and statistical features, such as part-of-speech (POS) information, rule-based knowledge, term frequencies, etc., can all be represented and integrated in this method
    - It was shown that very promising results can be obtained with this method

Named-Entity Extraction (7/10)

- Handling out-of-vocabulary (OOV) or unknown words
  - E.g., HMM
    - Divide the training data into two parts during training
      - In each half, every segment of terms or words that does not appear in the other half is marked as “Unknown”, such that the probabilities for both known and unknown words occurring in the respective named-entity states can be properly estimated
    - During testing, any segment of terms that is not seen before can thus be labeled “Unknown,” and the Viterbi algorithm can be carried out to give the desired results

Named-Entity Extraction (8/10)

- Handling out-of-vocabulary (OOV) or unknown words for spoken docs
  - Out-of-vocabulary (OOV) problem is raised due to the limitation in the vocabulary size of speech recognizer
  - OOV words will be misrecognized as other in-vocabulary words
    - Lose their true semantic meanings
- Tackle this problem using ASR & IR techniques
  - In ASR (automatic speech recognition)
    - Spoken docs are transcribed using a recognizer implemented with a lexical network modeling both word- and subword-level (phone or syllable) *n*-gram LM constraints
      - The speech portions corresponding to OOV words may be properly decoded into sequences of subword units

Named-Entity Extraction (9/10)

- Tackle this problem using ASR & IR techniques (cont.)
    - The subword *n*-gram LM is trained by the text segments corresponding to the low-frequency words not included in the vocabulary of the recognizer
  - In IR (Information Retrieval)
    - A retrieval process was performed using each spoken doc itself as a query to retrieve relevant docs from a temporal/topical homogeneous reference text collection
    - The indexing terms adopted here can be either word-level features, subword-level features, or both of them

Named-Entity Extraction (10/10)

- Tackle this problem using ASR & IR techniques (cont.)
  - Once the top-ranked text documents are selected, each decoded subword sequence within the spoken document, that are corresponding to a possible OOV word, can be used to match every possible text segments or word sequences within the top-ranked text documents
  - The text segment or word sequence within the top-ranked text docs that has the maximum combined score of phonetic similarity to the OOV word and relative frequency in the relevant text docs can thus be used to replace the decoded subword sequence of the spoken doc

$$\max_{w} \sum_{d \in D_r} P\left(e_{oov} \middle| w\right) \cdot P\left(w \middle| d\right) \cdot P\left(d \middle| q_s\right)$$

phone/syllable sequence of the OOV words

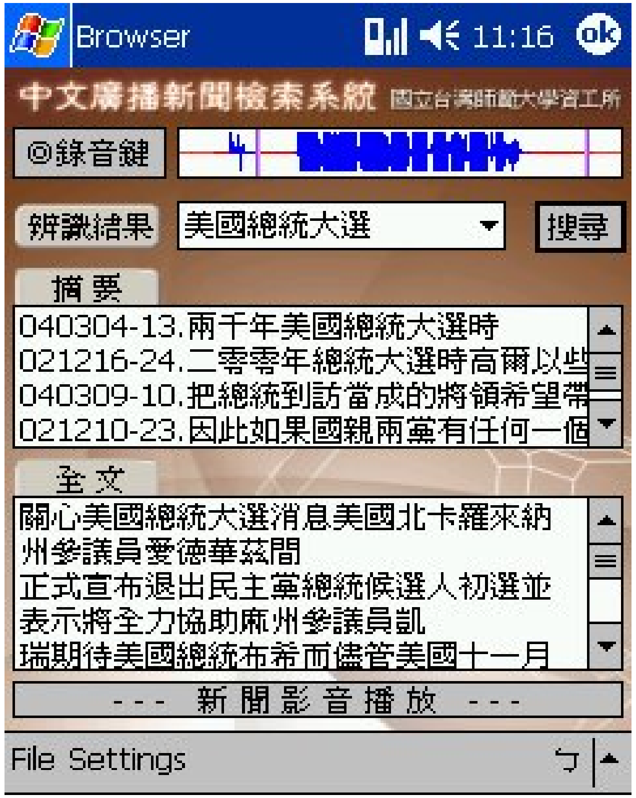
word in the top-ranked relevant text doc set

doc belonging to the top-ranked relevant text doc set

spoken doc

Prototype Systems Developed at NTNU (1/3)

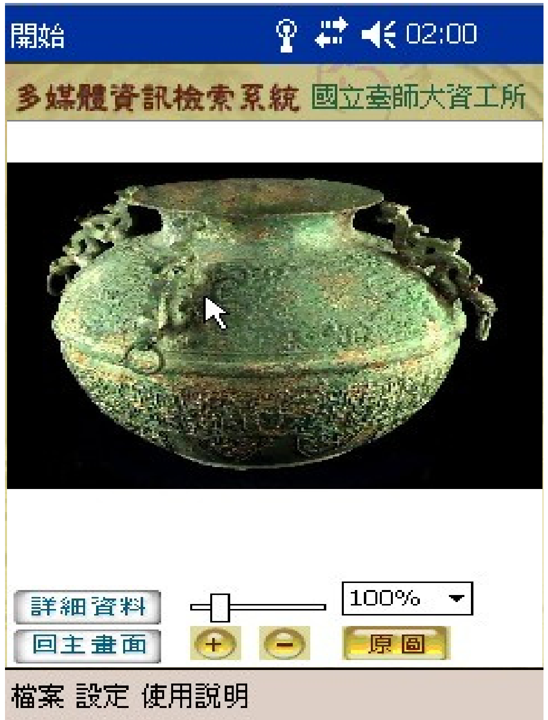
- Spoken Document Retrieval



http://sdr.csie.ntnu.edu.tw

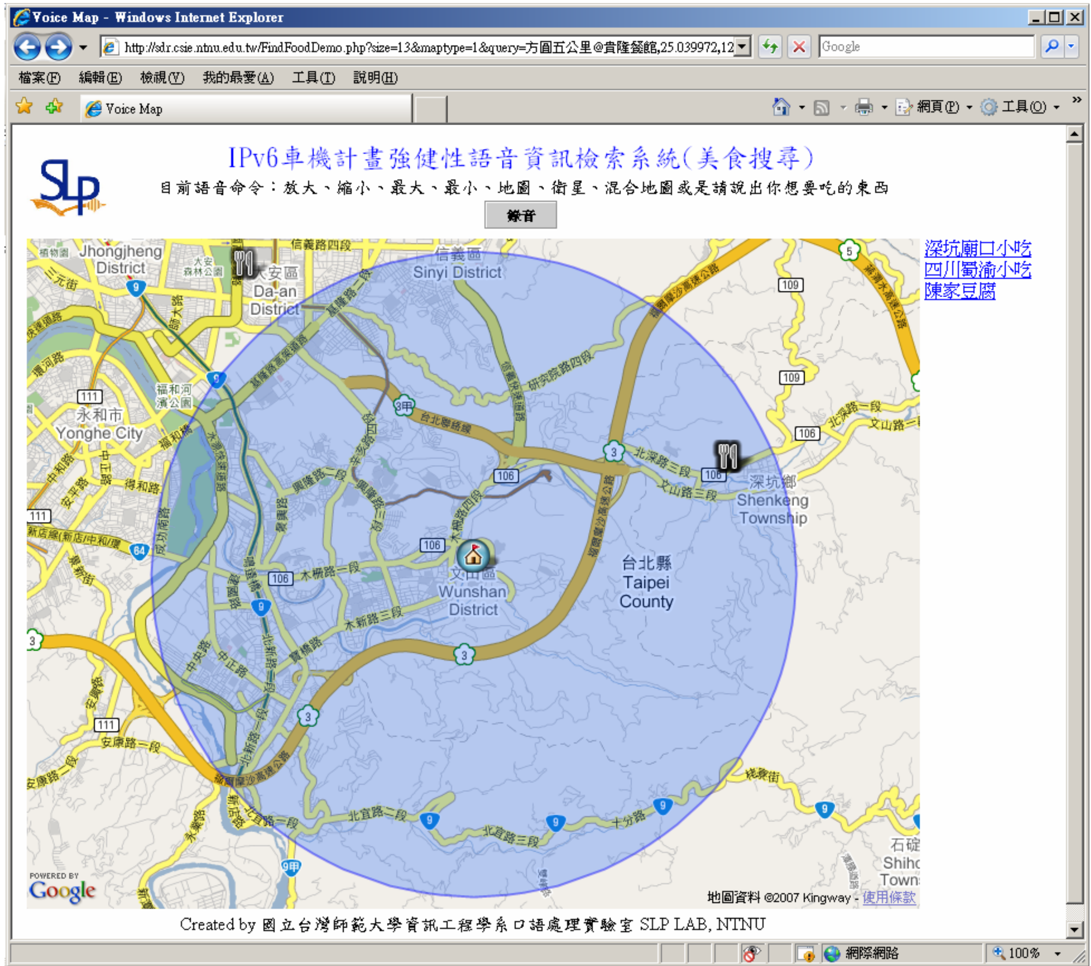
Prototype Systems Developed at NTNU (2/3)

- Speech Retrieval and Browsing of Digital Archives



Prototype Systems Developed at NTNU (3/3)

- Speech-based Information Retrieval for ITS systems



Query: 我想找方圓五公里賣臭豆腐的店家

– Projects supported by Taiwan Network Information Center (TWNIC)

Conclusions and Future Work

- Multimedia information access using speech will be very promising in the near future
  - Speech is the key for multimedia understanding and organization
  - Several task domains still remain challenging

- Spoken document retrieval (SDR) provides good assistance for companies in
  - Contact (Call)-center conservations: monitor agent conduct and customer satisfaction, increase service efficiency
  - Content-providing services such as MOD (Multimedia on Demand): provide a better way to retrieve and browse descried program contents

Thank You!

References

[R1] B. Chen et al., "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," *ACM Transactions on Asian Language Information Processing*, Vol. 3, No. 2, June 2004

[R2] J.R. Bellegarda, "Latent Semantic Mapping," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005

[R3] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005

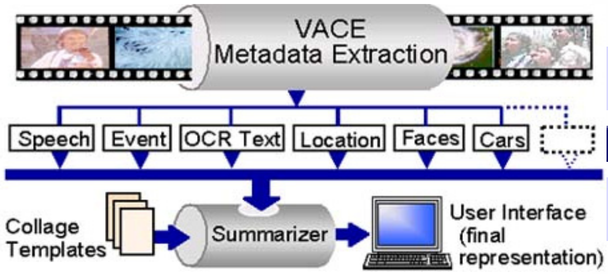
[R4] L.S. Lee and B. Chen, "Spoken Document Understanding and Organization," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005

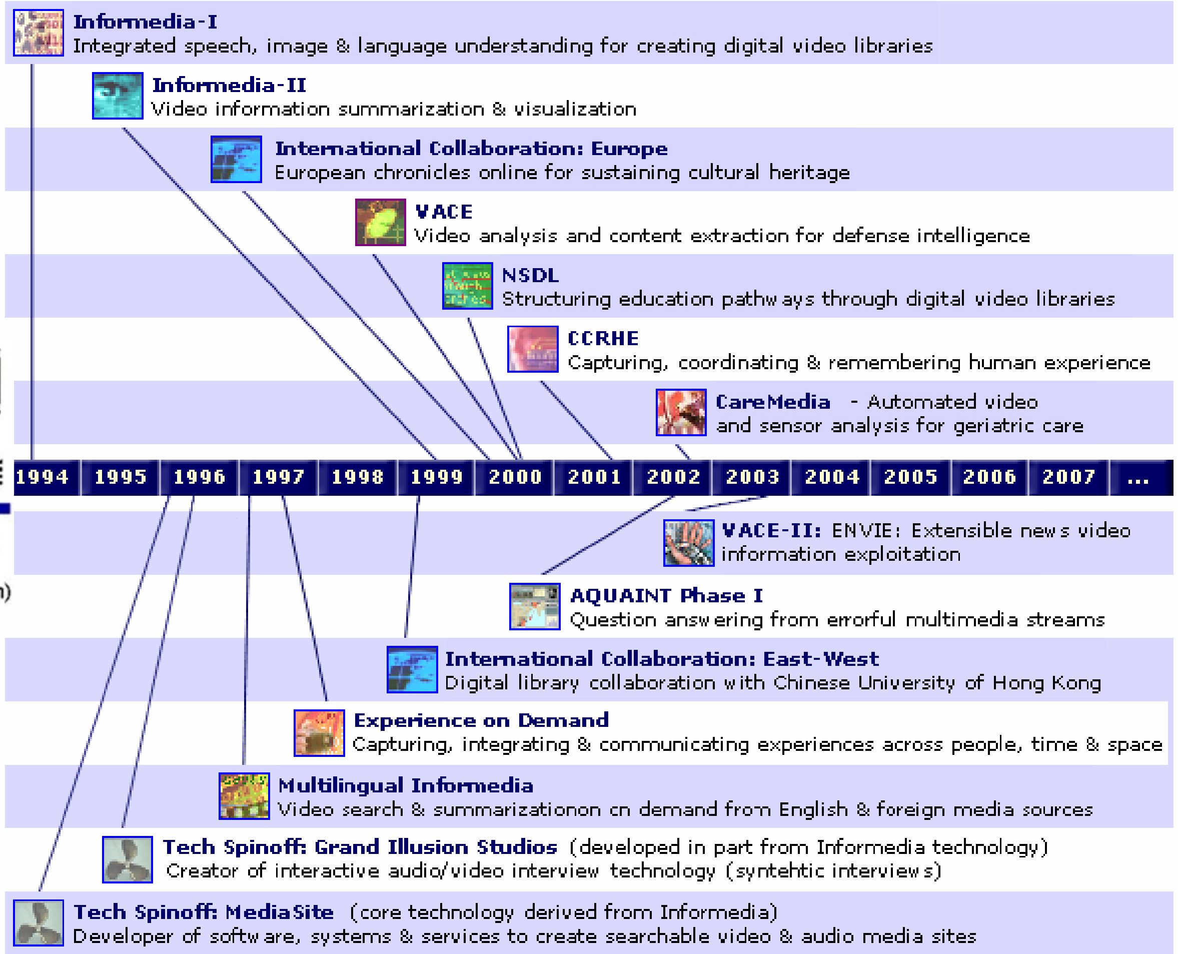
[R5] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning,* Vol. 42, 2001

[R6] B. Chen, "Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval," *Pattern Recognition Letters*, Vol. 27, No. 1, Jan. 2006

The Informedia System at CMU

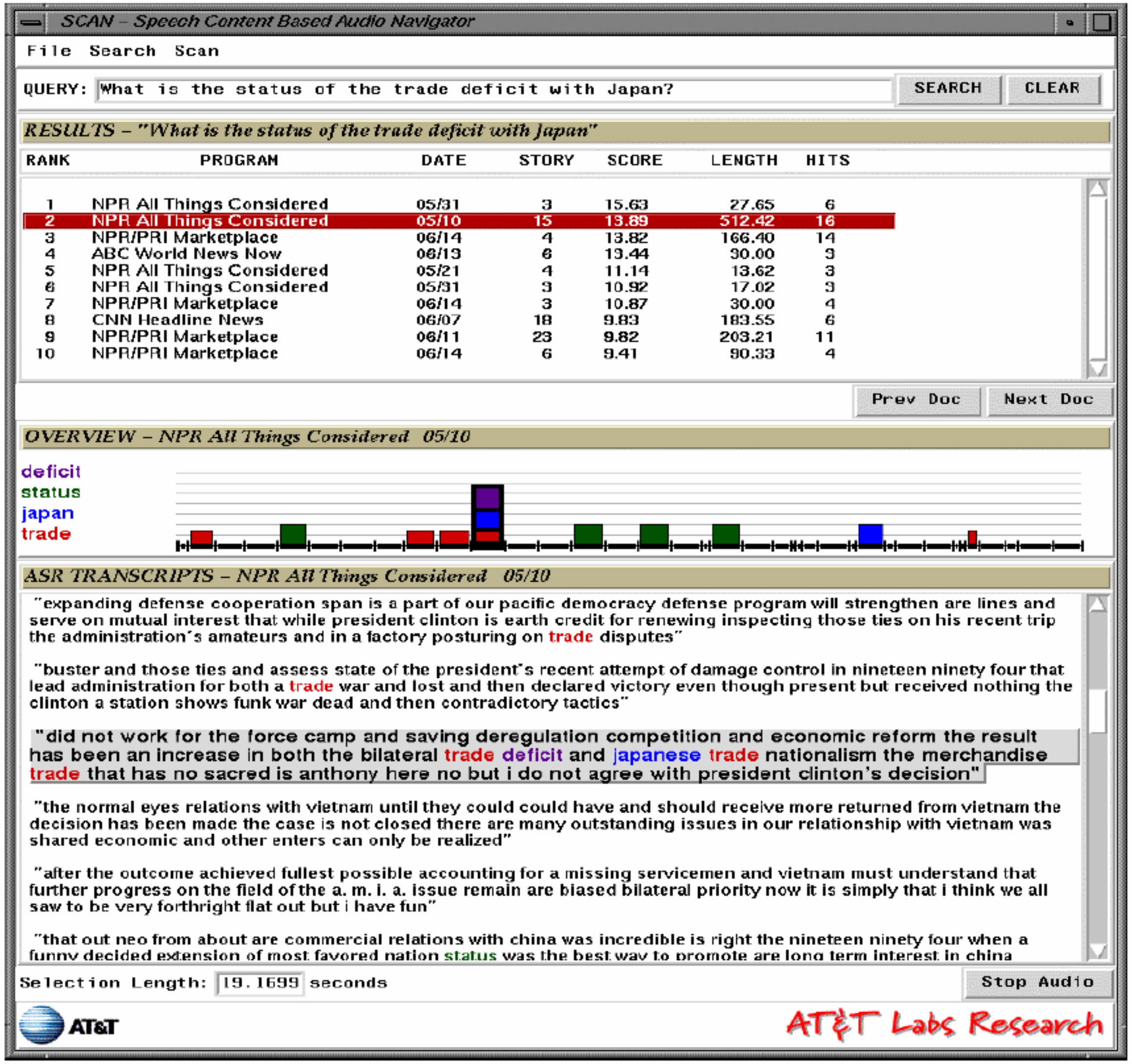
- Video Analysis and Content Extraction (VACE)
  - http://www.informedia.cs.cmu.edu/





AT&T *SCAN* System

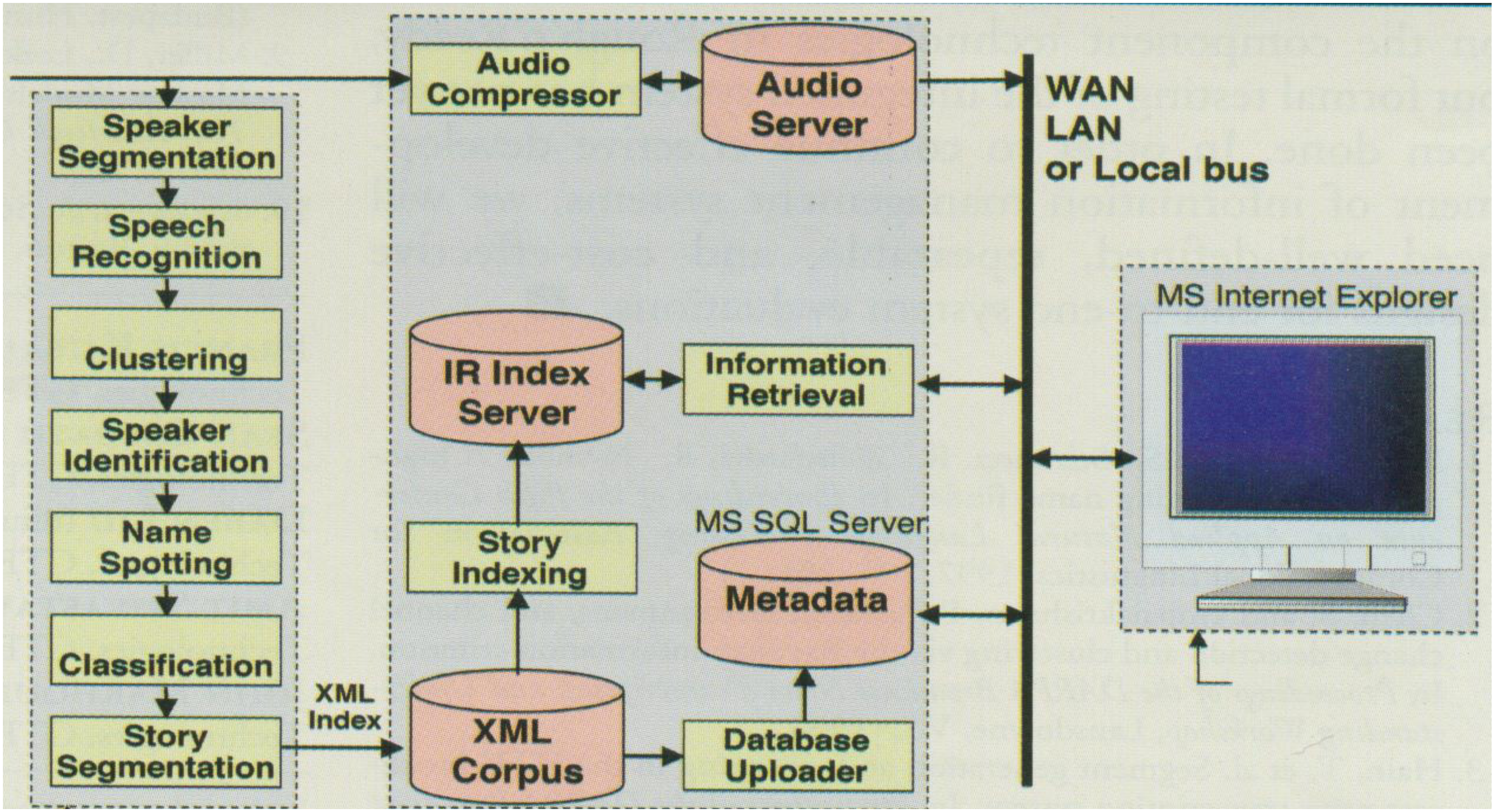
- SCAN: Speech Content Based Audio Navigator (1999)



Design and evaluate user interfaces to support retrieval from speech archives

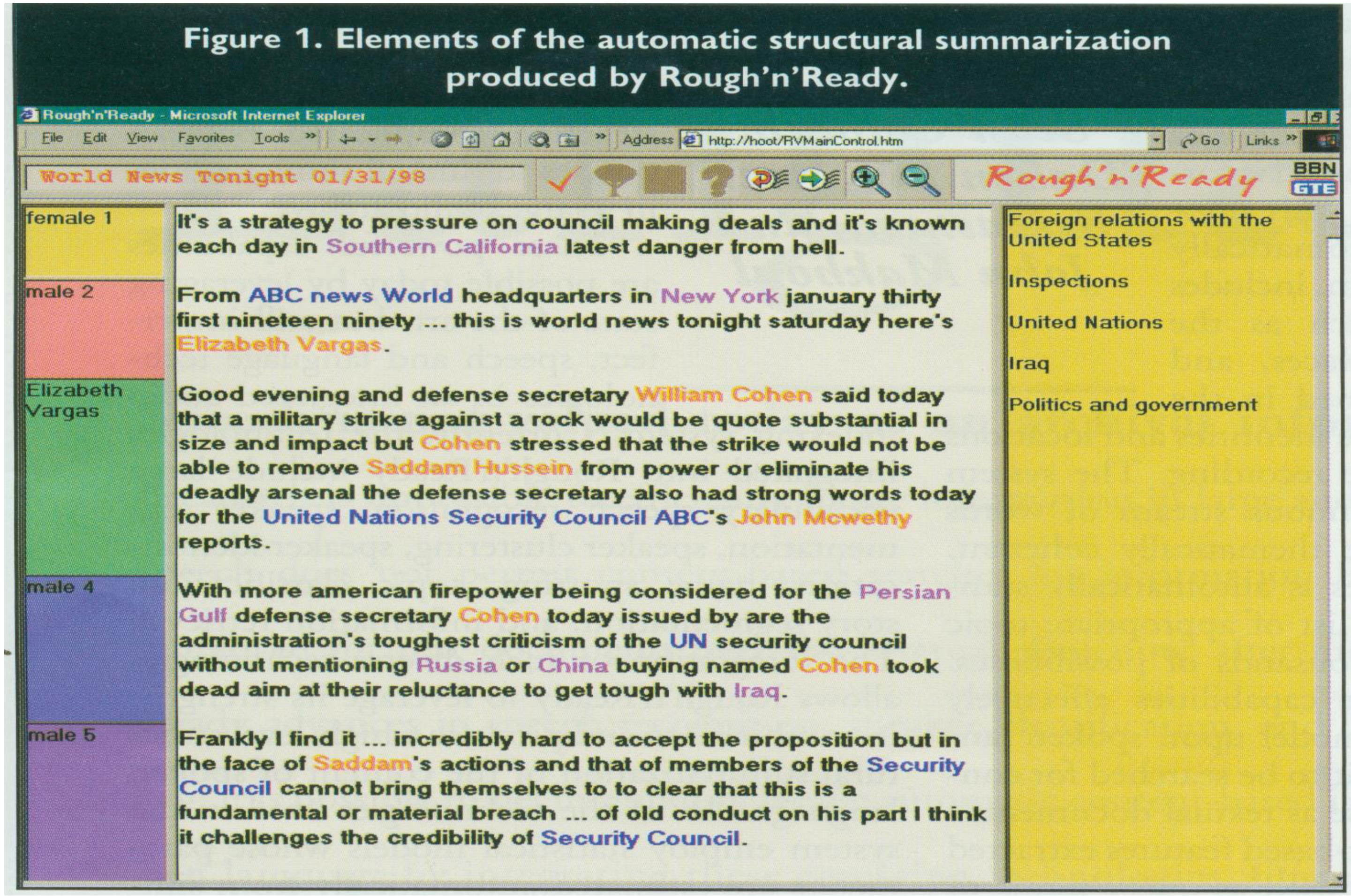
BBN *Rough'n'Ready* System (1/2)

- Distinguished Architecture for Audio Indexing and Retrieval (2002)



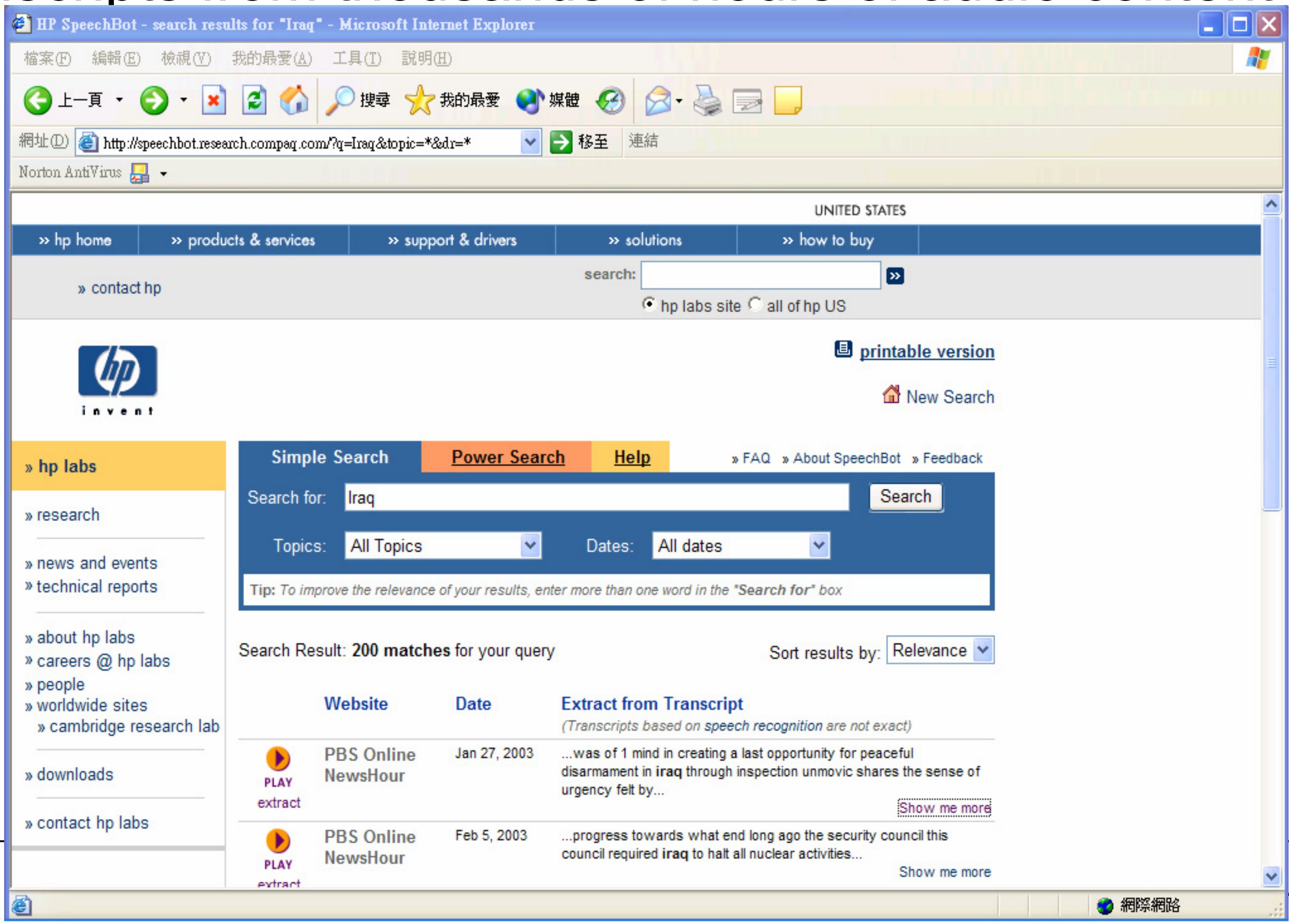
BBN *Rough'n'Ready* System (2/2)

- Automatic Structural Summarization for Broadcast News



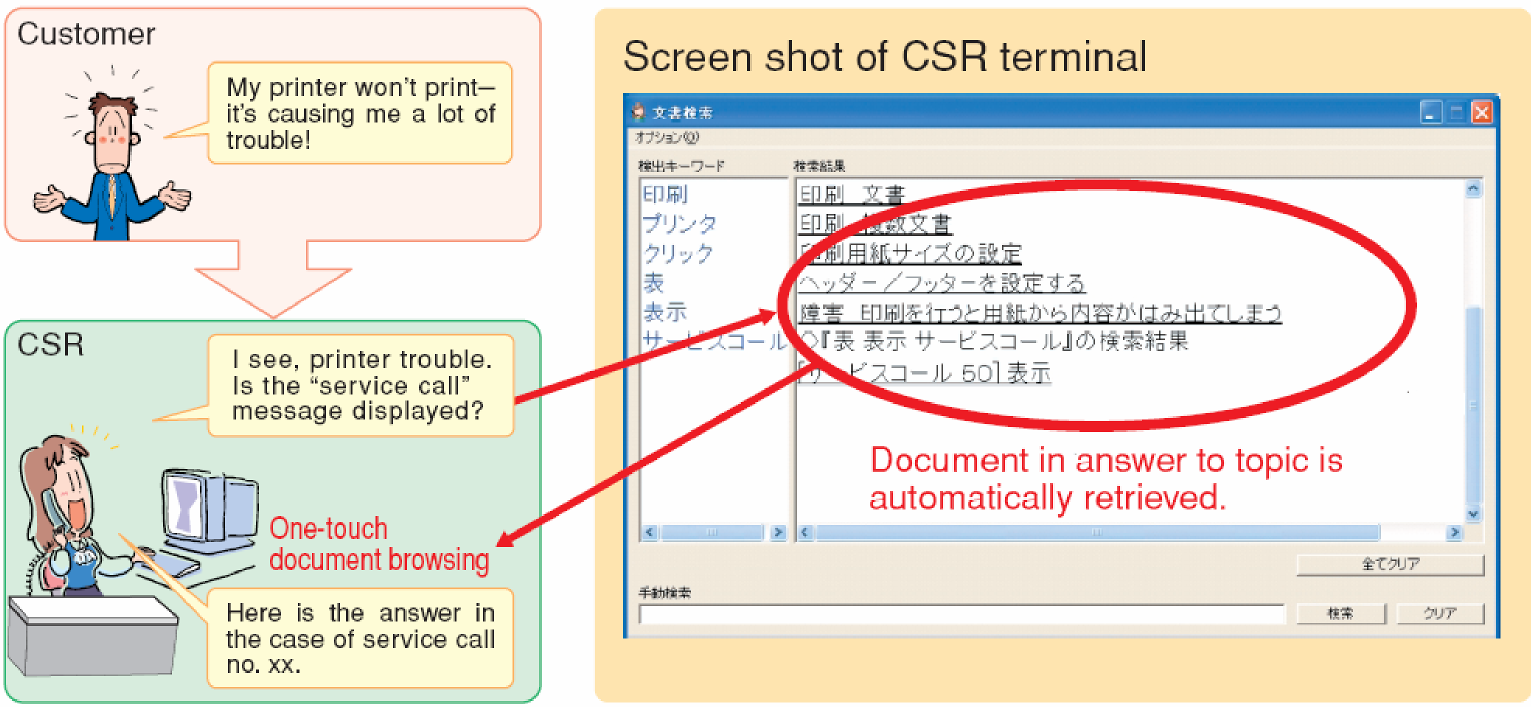
SpeechBot Audio/Video Search System at HP Labs

- An experimental web-based tool from HP Labs that used voice-recognition to create seachable keyword transcripts from thousands of hours of audio content



NTT Speech Communication Technology for Contact Centers

Automatic document-retrieval by speech recognition



– CSR: Customer Service Representative

Google Voice Local Search

http://www.google.com/goog411/