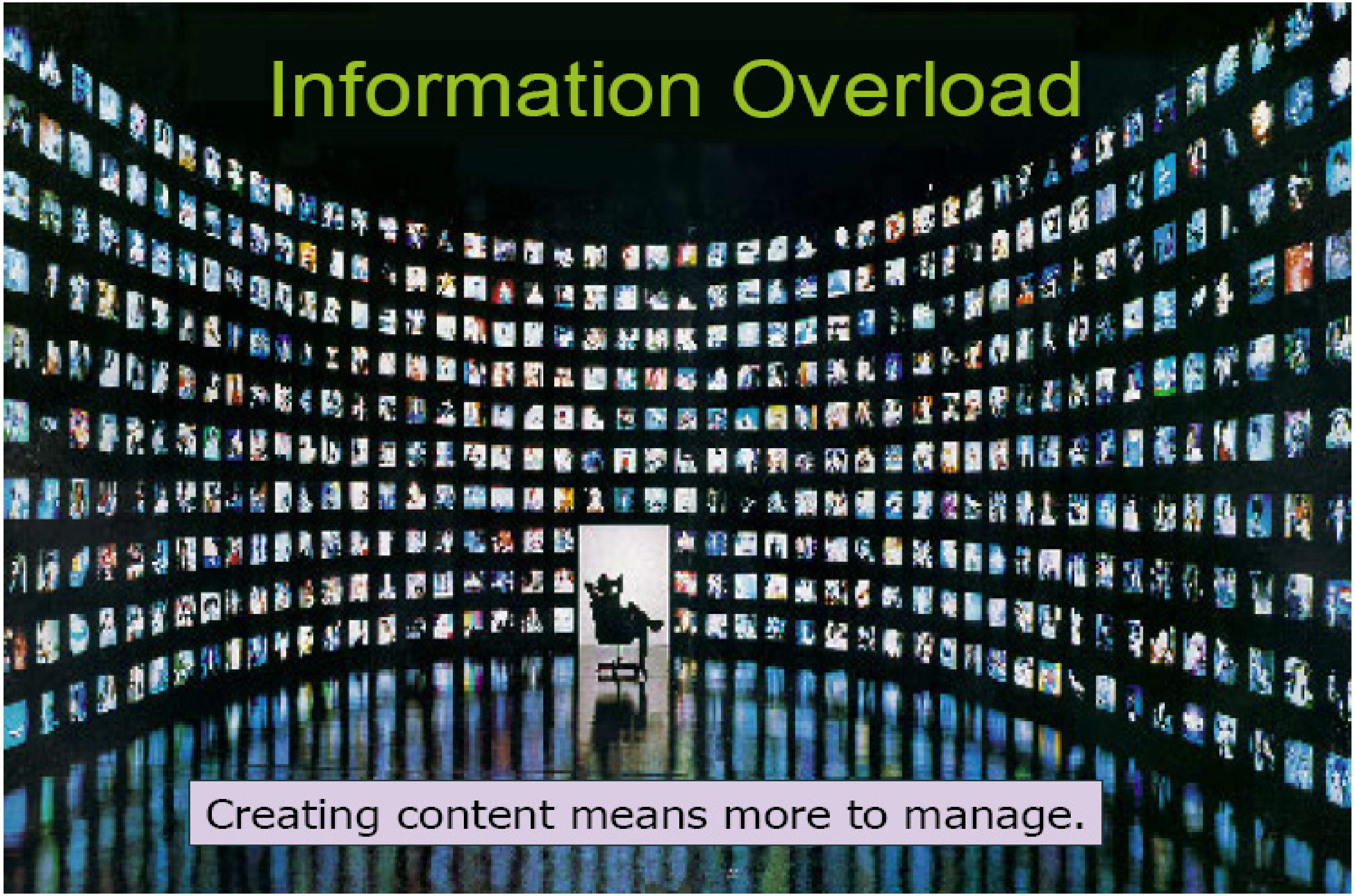
Relevance Language Modeling for Speech Recognition and Related Applications

Berlin Chen (陳柏琳)
Department of Computer Science & Information Engineering
National Taiwan Normal University
2011/12/28

Outline

- Introduction
- Automatic Speech Recognition (ASR)
- Relevance Language modeling for ASR
- Related Tasks: Speech Retrieval and Summarization
- Conclusions

Information -> Knowledge -> Wisdom ?



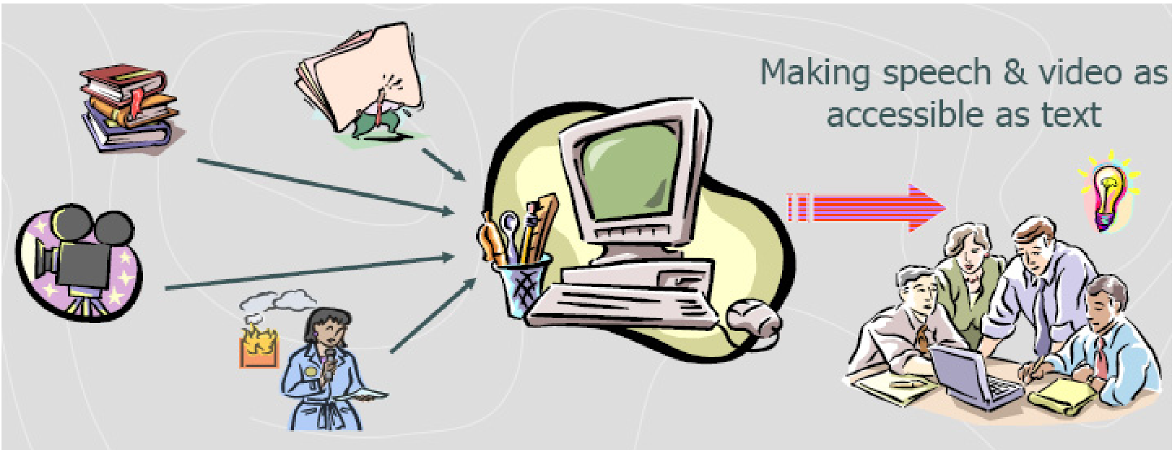
The figure is adapted from the presentation slides of Prof. Ostendorf at *Interspeech2009*

Introduction (1/2)

- Communication and search are by far the most popular activities in our daily lives
  - **Human-Computer Interaction:** Speech is the most nature and convenient means of communication between humans, and between humans and machines
    - A spoken language interface could be more convenient than a visual interface on a small device
    - Provide "anytime" and "anywhere" access to information
  - **Multimedia Content Processing:** Already over half of the internet traffic consists of video data
    - Though visual cues are important for search, the associated spoken documents often provide a rich set of semantic descriptions (e.g., transcripts, speakers, emotions, and scenes) for the data

Introduction (1/2)

- Automatic speech recognition (ASR)
  - Transcribe the linguistic contents of speech utterances
  - Play a vital role in multimedia information retrieval, summarization and mining, as well as computer-assisted language learning (CALL), such as
    - Transcribing spoken queries and documents
    - Determine pronunciation accuracy and intelligibility



The figure is adapted from the presentation slides of Prof. Ostendorf at *Interspeech 2009.*

Automatic Speech Recognition (ASR)

- Decision Rule of ASR (Risk-Minimization Principle)

$$W_{opt} = \arg\min_{W \in \mathbf{W}} Risk\left(W \middle| O\right)$$

$$= \arg\min_{W \in \mathbf{W}} \sum_{W' \in \mathbf{W}} Loss\left(W, W'\right) P\left(W' \middle| O\right)$$

Applying Bayes Theorem

$$\approx \arg\max_{W \in \mathbf{W}} P\left(W \middle| O\right)$$ Assumption of Using the "**0-1**" Loss Function

$$\approx \arg\max_{W \in \mathbf{W}} \frac{p(O \mid W) P(W)}{p(O)}$$

$$= \arg\max_{W \in \mathbf{W}} p(O \mid W) P(W)$$

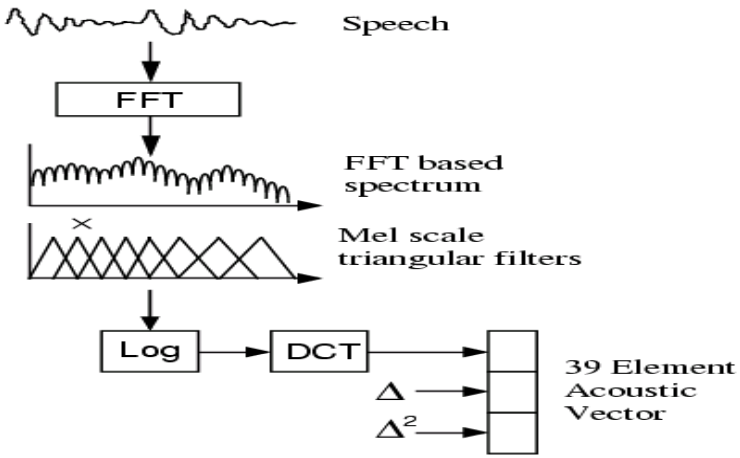
Linguistic Decoding

Feature Extraction & Acoustic Modeling   Language Modeling

  - The ASR problem is reduced to finding the most likely word sequence ***W*** in response to an input speech signal ***O***

Speech Feature Extraction

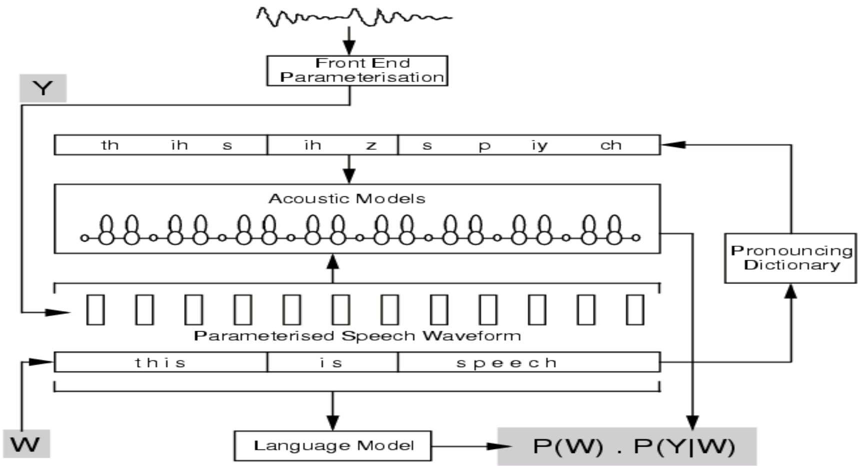
- The raw speech waveform is passed through feature extraction to generate relatively compact feature vectors at a frame rate of around 100 Hz
  - Parameterization: an acoustic speech feature is a simple compact representation of speech and can be modeled by cepstral features such as the Mel-frequency cepstral coefficient (MFCC)



raw (perception-driven) features vs. discriminant (posterior) features

Acoustic Modeling: HMMs (1/2)

- An inventory of phonetic hidden Markov models (HMMs) can constitute any given word in the pronunciation lexicon with two assumptions
  - **First-order (Markov) assumption**: the state transition depends only on the origin and destination
  - **Output-independent assumption**: all observation frames are dependent on the state that generated them, not on neighboring observation frames



Steve Young et al. The HTK Book. Version 3.4, March 2006

Acoustic Modeling: HMMs (2/2)

- Three fundamental problems

1. Computation of the probability (likelihood) of a sequence of observations given a specific HMM
   - Forward/backward algorithms for efficient computation
2. Determination of a best sequence of model states
   - Viterbi algorithm for state alignment
3. Adjustment of model parameters so as to best account for observed signals (or discrimination purposes)
   - Maximum Likelihood (ML), Maximum A Posteriori (MAP) and Discriminative Training (DT) criteria
     - DT considers not only the correct (or reference) transcript of a training utterance, but also the competing hypotheses for better model discrimination

M.J.F. Gales and S.J. Young. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing, 2008

Language Modeling: *n*-grams (1/2)

- For a word sequence $\boldsymbol{W}$, $P(\boldsymbol{W})$ can be decomposed into a product of conditional probabilities

chain (multiplication) rule

$$
\begin{aligned}
P(\boldsymbol{W}) &= P(w_1, w_2, \ldots, w_m) \\
&= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_m|w_1, w_2, \ldots, w_{m-1}) \\
&= P(w_1)\prod_{i=2}^{m} P(w_i|\underbrace{w_1, w_2, \ldots, w_{i-1}}_{\text{History of } w_i})
\end{aligned}
$$

- *n*-gram modeling: the history is put into V^{n-1} equivalence classes, where V is the vocabulary size

$$
P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx P(w_i|\underbrace{w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}}_{\text{History of length } n-1})
$$

  - Bigram (n=2) and trigram (n=3) are the most prevalent

$$
P(w_i|w_1, w_2, \ldots, w_{i-1}) \approx P(w_i|w_{i-2}, w_{i-1}) \ \text{ or } \ P(w_i|w_{i-3}, w_{i-2}, w_{i-1})
$$

R. Rosenfeld, "Two Decades of Statistical Language Modeling: Where Do We Go from Here?," Proceedings of IEEE, 2000

Language Modeling: *n*-grams (2/2)

- Known Weakness of *n*-grams
  - Sensitive to changes in the style or topic of the text on which they are trained
  - Assume the probability of next word in a sentence depends only on the identity of last *n*-1 words
    - Capture only local contextual information or lexical regularity of a language
- F. Jelinek said "*put language back into language modeling*"
  - Structure and topic models and language models have been proposed to harness extra information cues complementary to *n*-grams; e.g., a typical topic model

$$P_{\text{Topic}}\left(w_i \mid \text{History}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) \cdot P\left(T_k \mid \text{History}\right)$$

Steyvers and Griffiths, "Probabilistic topic models," in: *Handbook of Latent Semantic Analysis*, 2007.

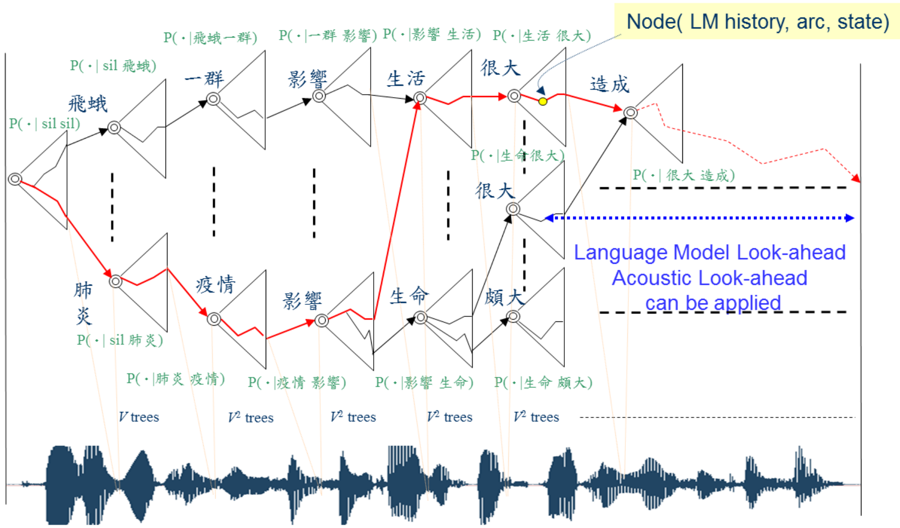
Linguistic Decoding (1/2)

- Find the most likely word sequence on top of the acoustic and language models and through
  - A dynamically-built word network: tree-copy search
  - A statically-built word network: finite state transducer, FST
- Efficient search algorithms and pruning techniques are highly demanded
  - Breadth-first search (BFS) with path pruning (beam search)
  - A* search (or stack decoding) with heuristics/evaluation functions
- Need to strike the balance between time and space requirements

Aubert, X. L., "An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition," Computer Speech and Language, 2002

Linguistic Decoding (2/2)

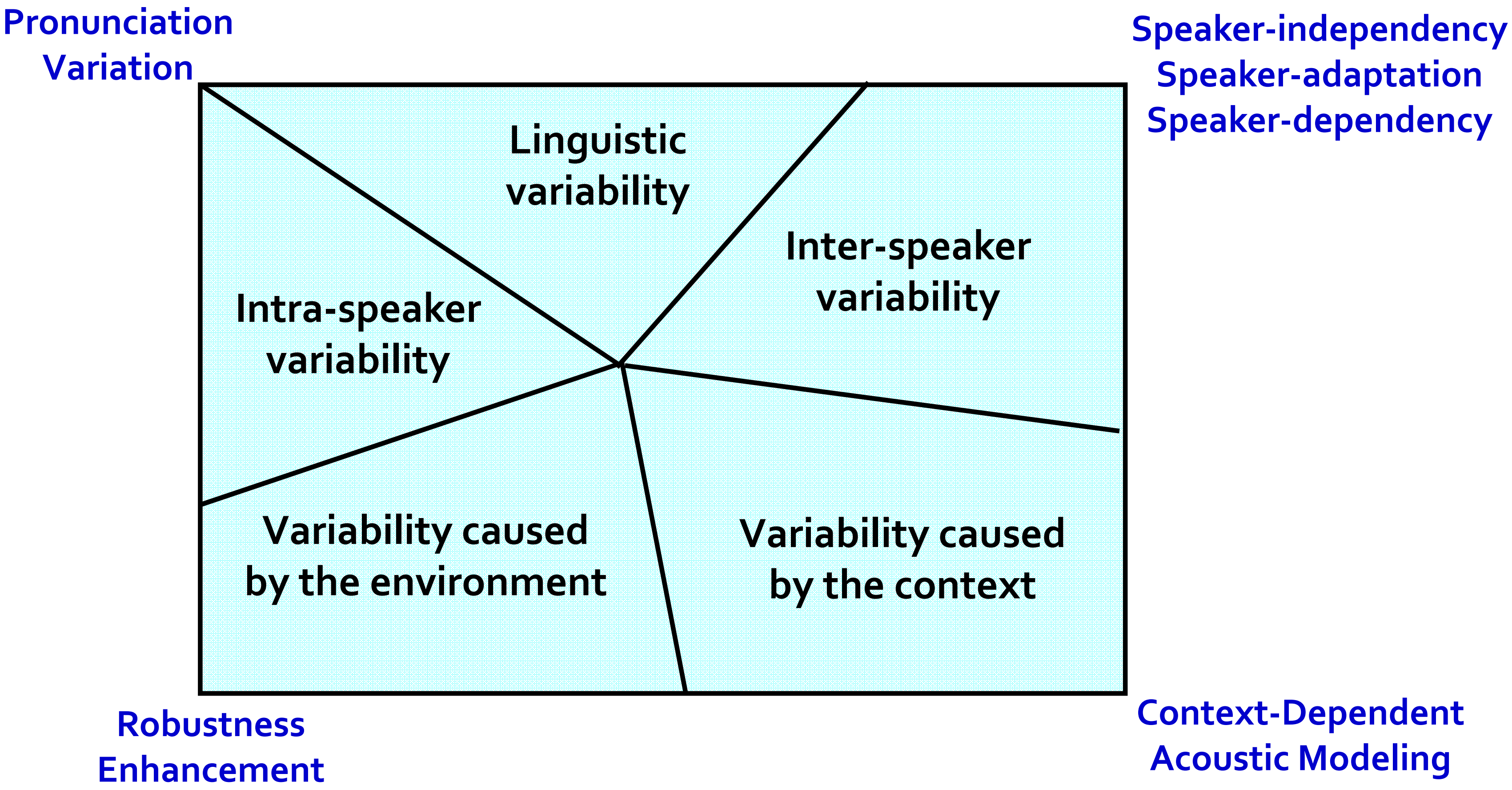
- E.g., tree-copy search with *n*-gram (bigram) models



  - The pronunciation lexicon is structured as a tree
  - Due to the constraints of *n*-gram language modeling, a word's occurrence is dependent on the previous *n*-1 words
  - We have to search through all possible tree copies from the start time to the end time of the utterance to find a best sequence of word hypotheses

ASR Robustness is Crucial

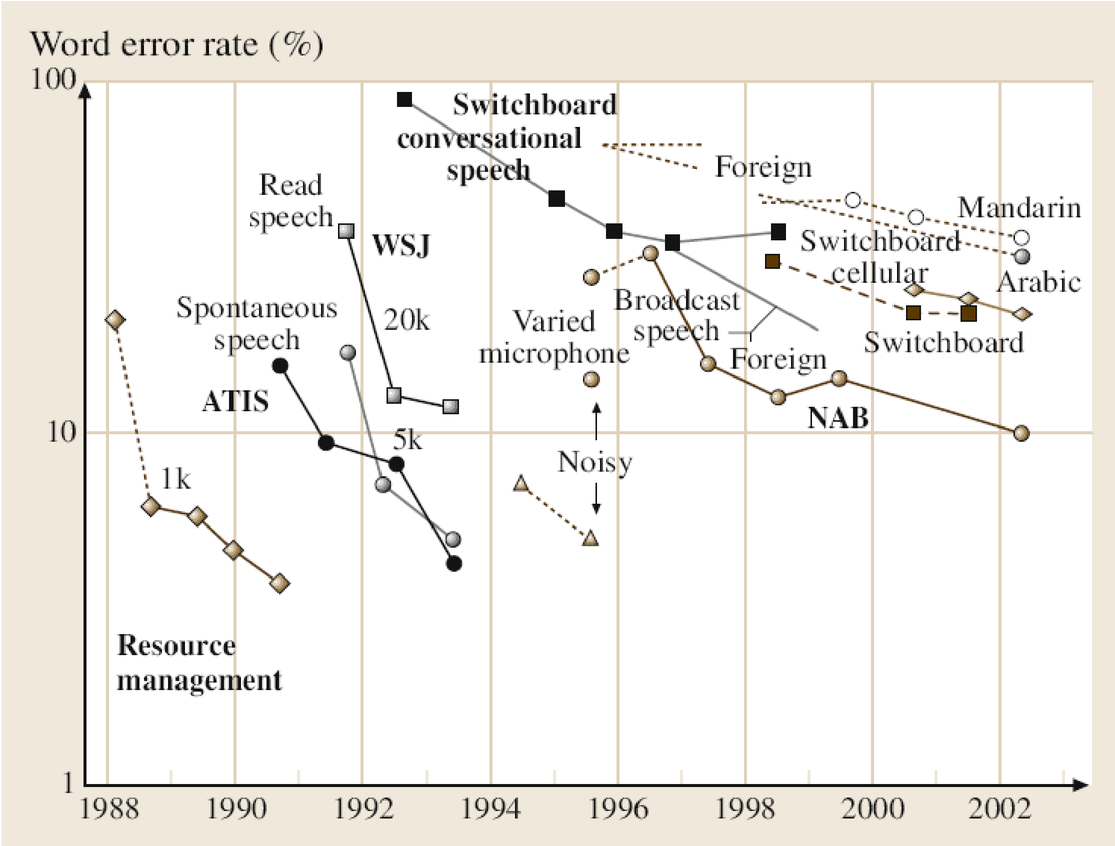
- The difficulty of ASR is further exacerbated by the speaker and environment variability



C.-H. Lee et al. (eds.) Automatic Speech and Speaker Recognition: Advanced Topics , 1996.

State-of-the-art ASR Performance

- Word error rate (WER) performance over time for a range of DARPA large-vocabulary speech recognition tasks



J. Benesty et al. (ed.) *Springer Handbook of Speech Processing*, Springer, 2007

Applications of ASR

- Multimedia (spoken document) retrieval and organization
  - Speech-driven Interface and multimedia content processing
  - Work in association with information retrieval techniques
  - A wild variety of potential applications (to be introduced later)
- Computer-Aided Language Learning (CALL)
  - Speech-driven Interface and multimedia content processing, in in conjunction with natural language processing techniques
    - Synchronization of audio/video learning materials
    - Automatic pronunciation assessment/scoring
    - Automated reading tutor
- Among many others

IEEE Signal Processing Magazine 25(3), 2008 (*Spoken Language Technology*)
IEEE Signal Processing Magazine 22(5), 2005 (*Speech Technology and Systems in Human-Machine Communication*)

Prototype and Deployed Systems

- *Informedia* System at Carnegie Mellon Univ.
- *Rough'n'Ready* System at BBN Technologies
- SpeechBot Audio/Video Search System at HP Labs
- IBM Speech Search for Call-Center Conversations & Call-Routing, Voicemails, Monitoring Global Video and Web News Sources (*TALES*)
- Google's *411 Voice Search*
- MIT Lecture Browser
- Apple's *Siri*

We are witnessing the golden age of ASR!

http://www.apple.com/iphone/features/siri.html

Relevance Language Modeling for ASR (1/4)

- Investigate a novel use of relevance information cues to dynamically complement (or adapt) the conventional *n*-gram models, assuming that
  - During ASR, a search history $H = h_1, h_2, \ldots, h_L$ is a sample from a relevance class R describing some semantic content
  - A probable word w that immediately succeeds the H is a sample from R

$$H : h_1, h_2, \ldots, h_L \xrightarrow{P(w|H)} w$$

search history

word being predicted

Relevance Class *R*

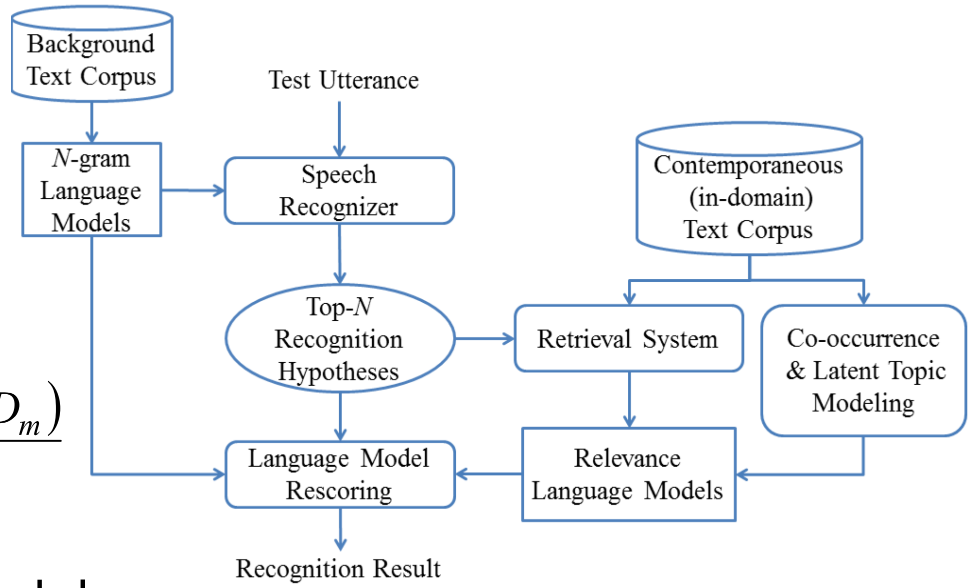
How to represent the relevance class *R* ?

K.-Y. Chen, B. Chen, "Relevance language modeling for speech recognition," *ICASSP2011.*

Relevance Language Modeling for ASR (2/4)

- Leverage the top-*M* relevant documents of the search history to approximate the relevance class R
  - Take H as a query to retrieve relevant documents
  - **R**elevance **M**odel: Multinomial view (*bag-of-words modeling*) of R

$$P_{\mathrm{RM}}(w|H)=\frac{P_{\mathrm{RM}}(H,w)}{P_{\mathrm{RM}}(H)}$$
$$=\frac{\sum_{m=1}^{M}P(D_m)P(H,w\mid D_m)}{\sum_{m=1}^{M}P(D_m)P(H\mid D_m)}$$
$$=\frac{\sum_{m=1}^{M}P(D_m)P(w\mid D_m)\prod_{l=1}^{L}P(h_l\mid D_m)}{\sum_{m=1}^{M}P(D_m)\prod_{l=1}^{L}P(h_l\mid D_m)}$$



  - Combined with *n*-gram models

$$P_{\mathrm{Adapt}}(w|H)=\lambda\cdot P_{\mathrm{RM}}(w|H)+(1-\lambda)\cdot P_{\mathrm{BG}}(w|h_{L-1},h_L)$$

Relevance Language Modeling for ASR (3/4)

- Further incorporation of latent topic information
  - A shared set of latent topic variables $\{T_1, T_2, \ldots, T_K\}$ is used to describe "*word-document*" co-occurrence characteristics

$$P(w \mid D_m) = \sum_{k=1}^{K} P(w \mid T_k) P(T_k \mid D_m)$$

$$P_{\mathrm{TRM}}(H, w) = \sum_{m=1}^{M} \sum_{k=1}^{K} P(D_m) P(T_k \mid D_m) P(w \mid T_k) \prod_{l=1}^{L} P(h_l \mid T_k)$$

- Alternative modeling of pairwise word associations

$$P_{\mathrm{PRM}}(h_l, w) = \sum_{m=1}^{M} P(D_m) P(h_l \mid D_m) P(w \mid D_m)$$

$$P_{\mathrm{PRM}}(w \mid H) = \sum_{l=1}^{L} \alpha_l \cdot P_{\mathrm{PRM}}(w \mid h_l)$$

$$P_{\mathrm{TPRM}}(h_l, w) = \sum_{m=1}^{M} \sum_{k=1}^{K} P(D_m) P(T_k \mid D_m) P(h_l \mid T_k) P(w \mid T_k)$$

N T N U

Relevance Language Modeling for ASR (4/4)

- Tested on a large vocabulary broadcast new recognition task
  - Character error rate (CER) results (the lower the better)

n-gram	RM	TRM	PRM	TPRM	PLSA	LDA	Cache	TBLM
20.08	19.29	19.08	19.23	19.09	19.15	19.15	19.86	20.02

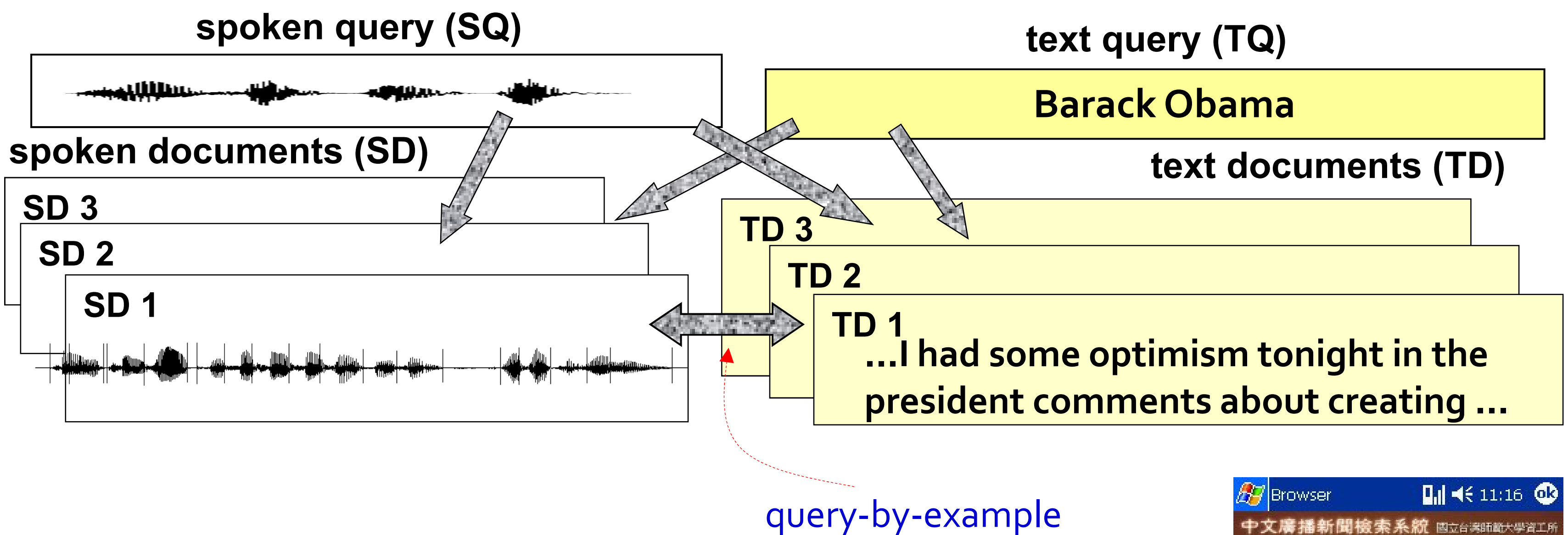
  - The various RM models achieve results compared to PLSA and LDA (topic models) and are considerably better than Cache and TBLM (trigger-based language model)
  - The various RM models are more efficient than PLSA and LDA
    - The various RM probabilities can be easily composed on the basis of the component probability distributions that were trained beforehand, without recourse to any complex inference procedure during the recognition (or rescoring) process
      - Computationally tractable and feasible for ASR

Speech Retrieval

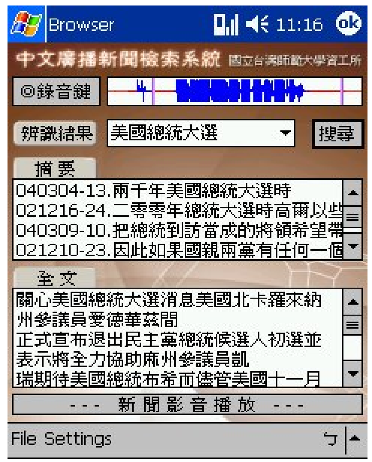
- Robustly Index spoken documents with speech recognition techniques
  - Explore better ways to represent the recognition hypotheses of spoken documents beyond the top scoring ones
  - Hybrid of words and subwords (phone/syllable/character *n*-grams) for indexing
- Retrieve relevant spoken documents in response to a user query
  - Spoken Document Retrieval (SDR)
    - Find spoken documents that are "topically related" to a given query
  - Spoken Term Detection (STD)
    - Find "literally matched" spoken documents where all/most query terms should be present (much like Web search)

Scenarios of Spoken Document Retrieval (SDR)

- Scenarios



  - SQ/SD is the most difficult
  - TQ/SD is studied most of the time
    - "query-by-example": e.g., use text news documents to retrieve relevant broadcast news documents
      - Useful for news monitoring and tracking

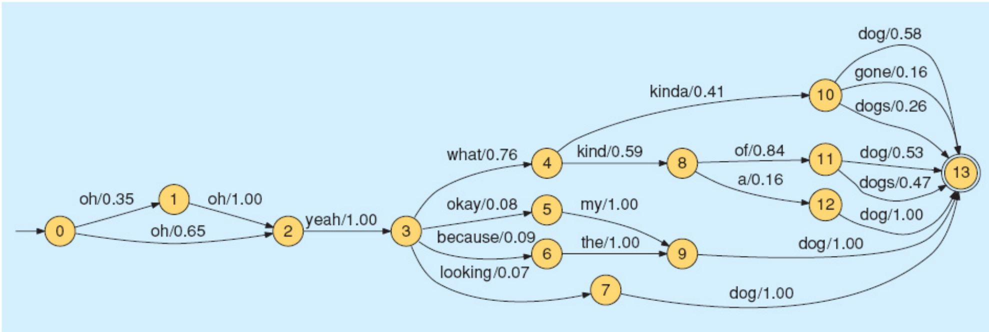


B. Chen et al., "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," *Interspeech 2005*

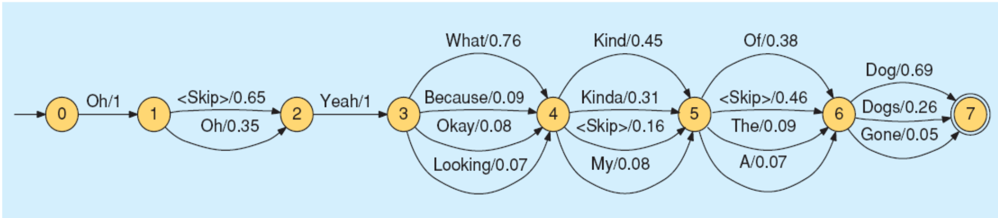
Representations of Spoken Queries and Documents

- Lattice/confusion network structures for retaining multiple recognition hypotheses

Lattice



Confusion Network



Position-Specific Posterior Probability Lattices

0	1	2	3	4	5	6	7
Oh 1.0	**Yeah** .65	**What** .46	**Kind** .27	Dog .26	EOS .34	**EOS** .44	EOS .16
—	Oh .35	Yeah .35	What .27	**Of** .23	**Dog** .29	Dog .09	—
	—	Because .06	Kinda .19	Kind .16	Dogs .13	Dogs .06	
		Okay .05	The .06	Kinda .11	Of .13	—	
		Looking .05	My .05	Dogs .05	A .03		
		—	Dog .05	EOS .05	Gone .02		
			....	....	—		

C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine* , 2008

Retrieval Models for SDR

- Information retrieval (IR) models, for example, 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

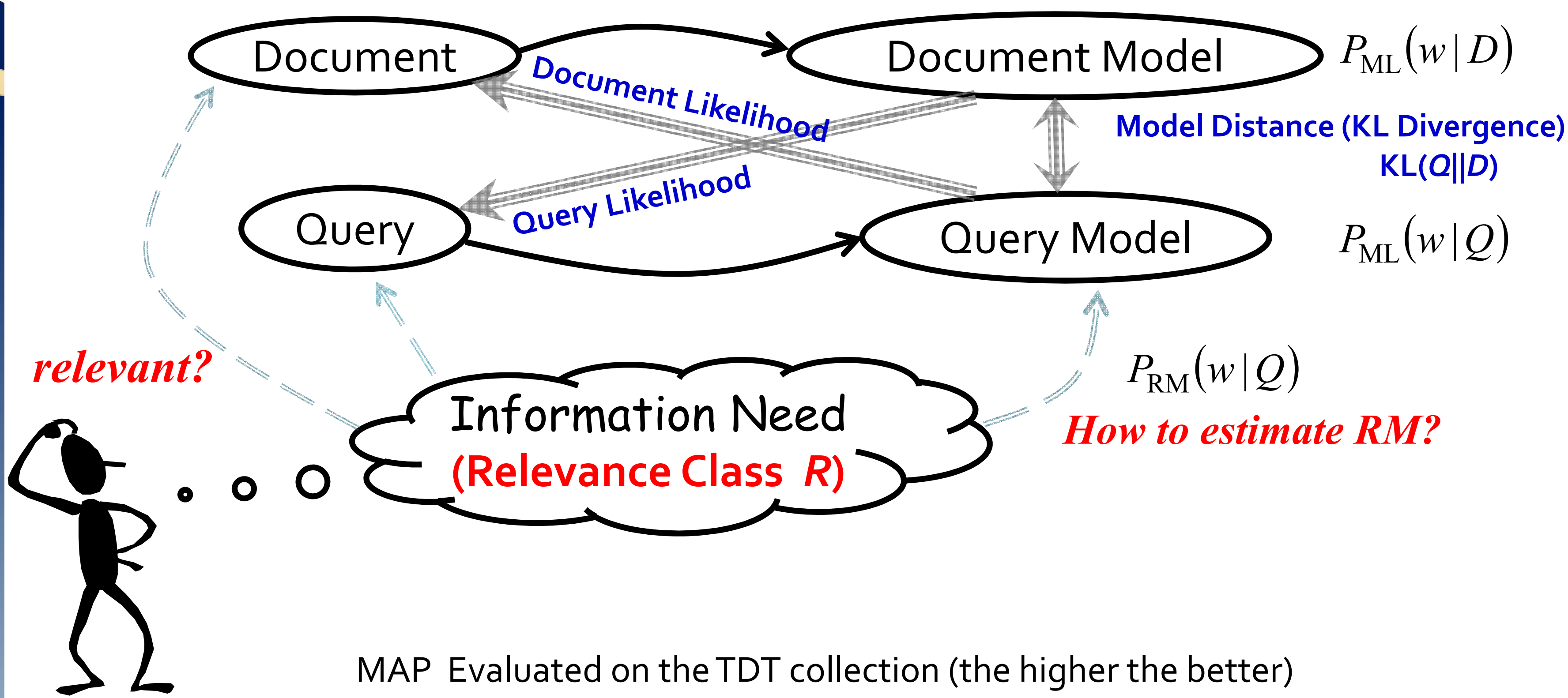
relevant ?

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二十二期在內的兩百架蘇霍伊戰鬥機。

B. Chen, "Word topic models for spoken document retrieval and transcription," *ACM Transactions on Asian Language Information Processing*, 2009

Relevance Language Modeling for SDR

- Schematic illustration

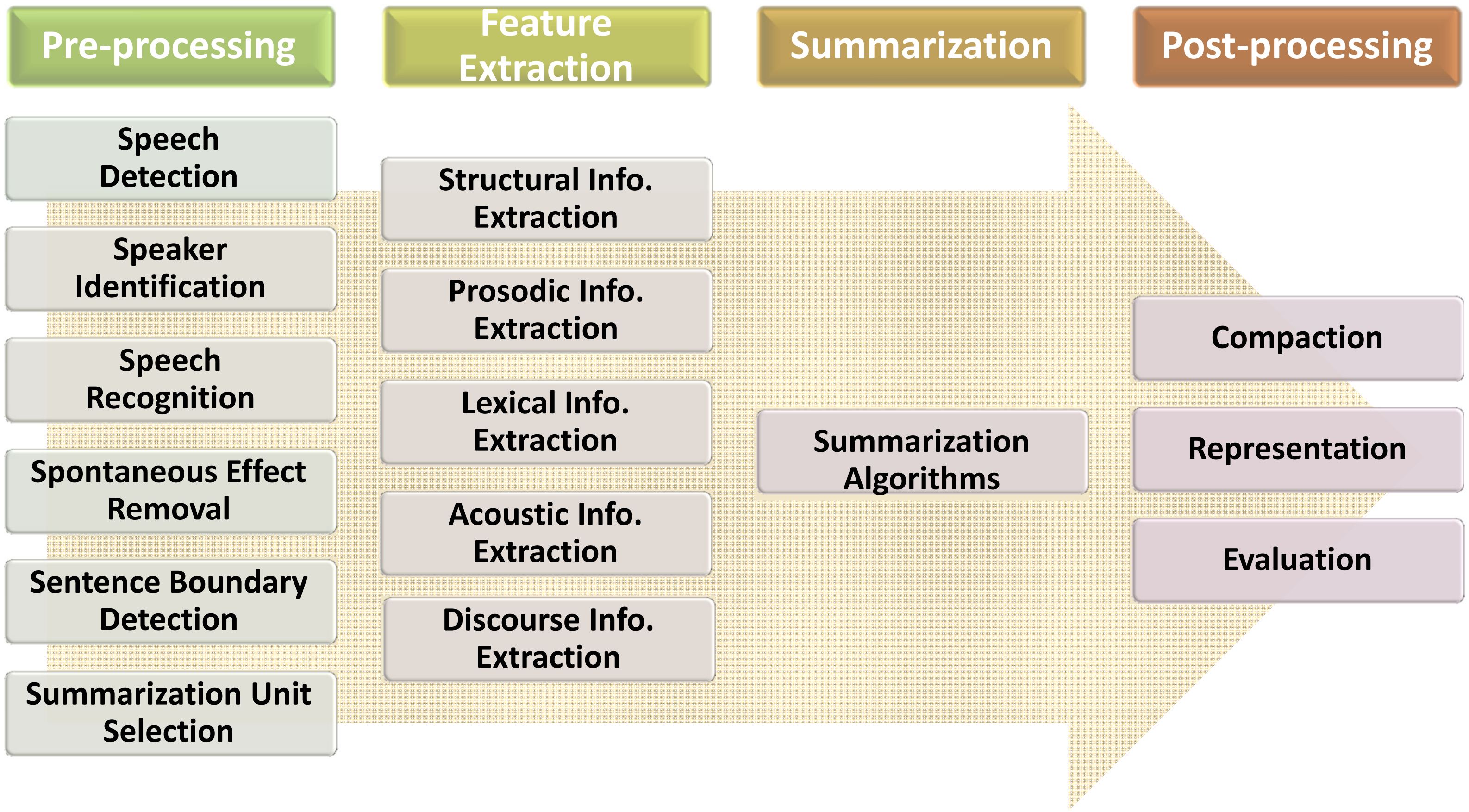


MAP Evaluated on the TDT collection (the higher the better)

ULM	RM	TRM	RM+NR	TRM+NR	PLSA	LDA
0.323	0.364	0.394	0.392	0.402	0.345	0.341

B. Chen et al., "Query modeling for spoken document retrieval," *ASRU2011.*

Extractive Speech Summarization

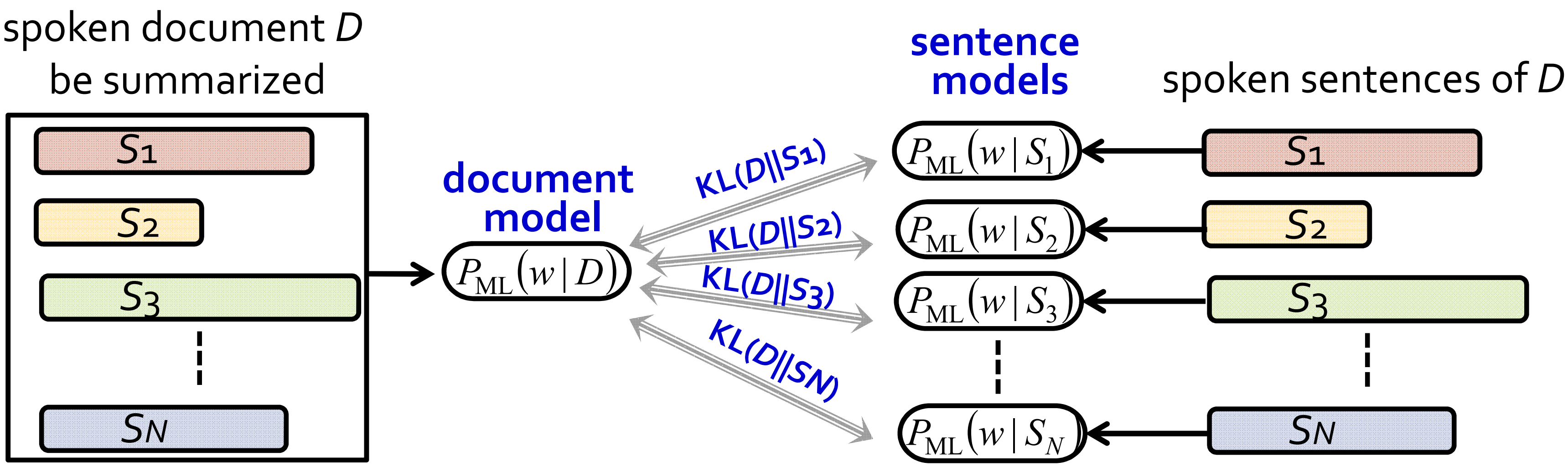


B. Chen and S.-H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 2012.

B. Chen et al., "Extractive speech summarization using evaluation metric-related training criteria," to appear in *Information Processing & Management*, 2012.

Relevance Language Modeling for Summarization

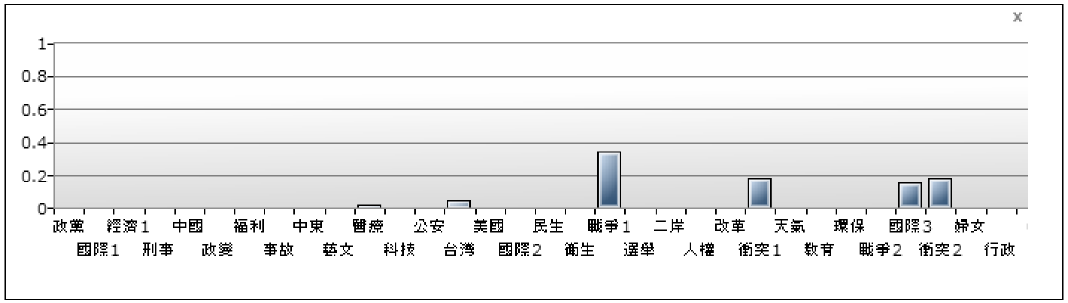
- Schematic illustration



$$S^* = \arg\min_{S_n} \lambda \, \mathrm{KL}(D \| S_n) - (1-\lambda)\, \mathrm{KL}(\mathbf{S} \| S_n)$$

  - Iteratively select important sentences *Sn* that have a small model distance to *D* but have a large distance to the set ***S*** of already selected sentences
  - Leverage sentence-specific **relevance model** (RM) and **non-relevance model** (NR) to enhance each sentence model

NTNU Lecture/News Browsing System



Conclusions

- Multimedia information access (over the Web) 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
- Speech retrieval and summarization provide good assistance for companies, for instance, 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
- Speech processing technologies are expected to play an essential role in computer-aided (language) learning