Speech Recognition and Spoken Document Retrieval for Mandarin Chinese

Hsin-min Wang

Institute of Information Science, Academia Sinica, Taiwan
Email: whm@iis.sinica.edu.tw

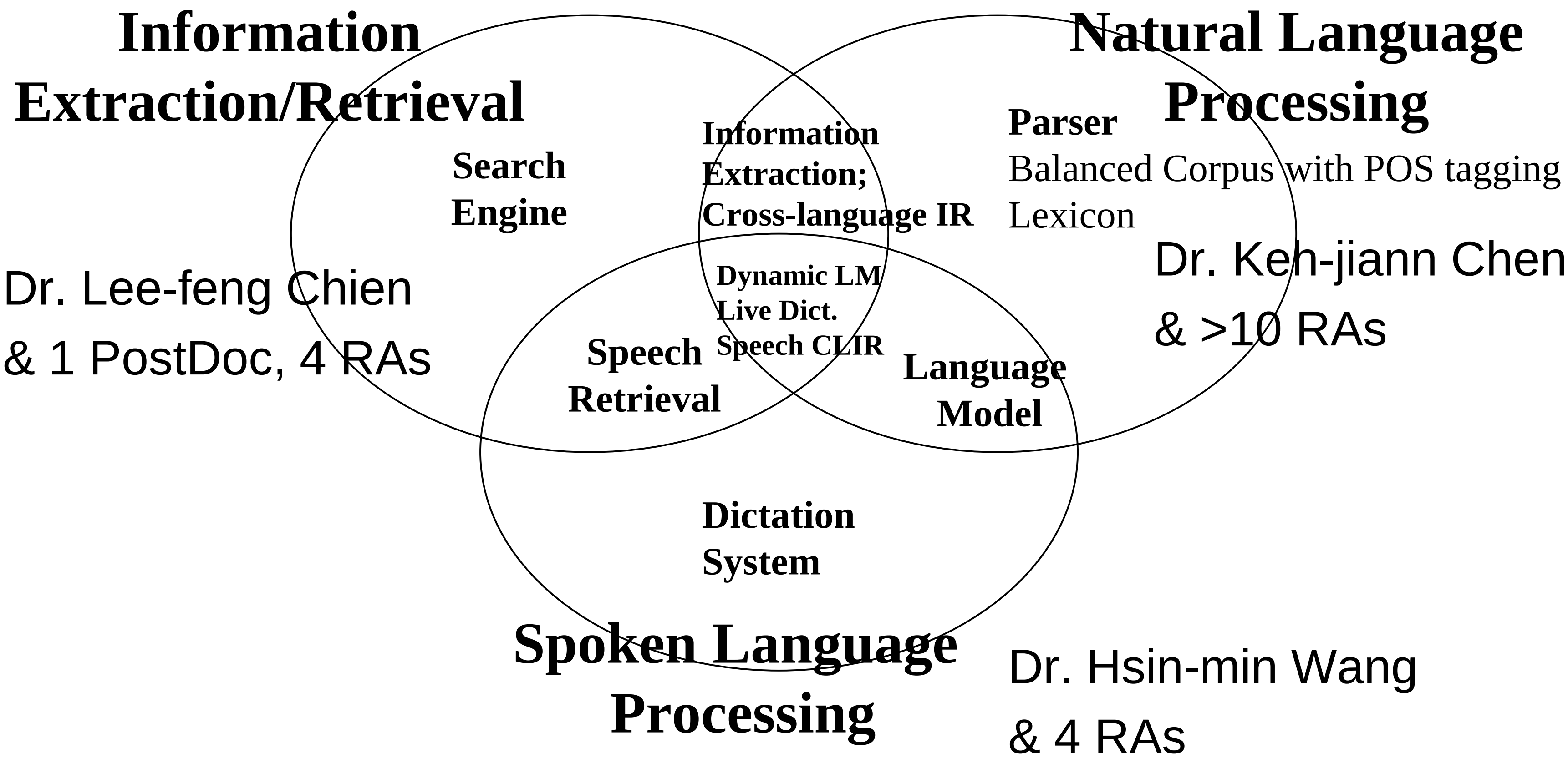
3rd SINO FRANCO WORKSHOP 2002/3/26-28

Outline

- Introduction to the Chinese Information Processing Laboratory
- Mandarin Chinese Large-Vocabulary Continuous Speech Recognition (LVCSR)
- Mandarin Chinese Spoken Document Retrieval (SDR)
  - Multi-scale overlapping N-gram indexing
  - Vector-space-based model
  - HMM/N-gram-based model
- Conclusion

Introduction to The Chinese Information Processing Laboratory

Research Paradigm of the Chinese Information Processing Lab



Mandarin Chinese Large-Vocabulary Continuous Speech Recognition (LVCSR)

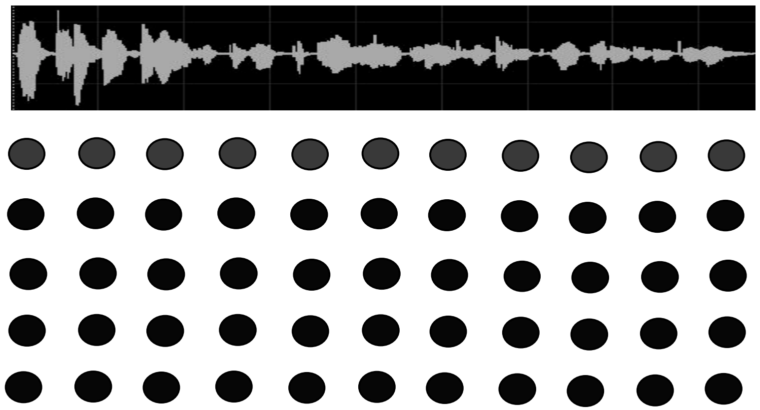
Characteristic of Mandarin Chinese

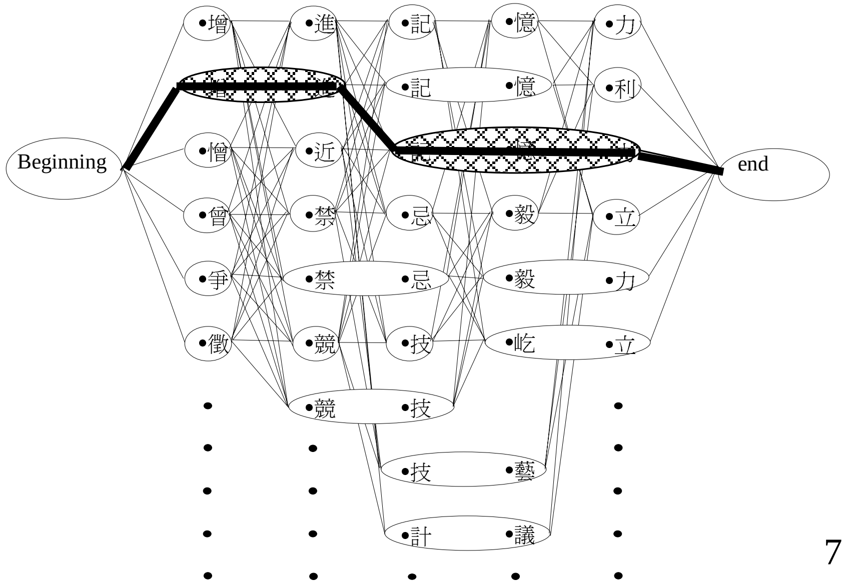
- 400 syllables
  - full phonological coverage in Mandarin Chinese
- 13,000 (Big5-coded, traditional) characters
  - full textual coverage in written Chinese
  - each character pronounced as a syllable
  - 6,800 GB-coded simplified Chinese characters
- Unknown number of Chinese words
  - one to several characters per word
  - character combinations create different meanings – new (or unknown) words
  - a foreign word may be translated into different Chinese words , e.g. Kosovo: 科索沃, 科索佛, 科索夫, 科索伏, 柯索佛

Multiple-pass Search for Speech Recognition

$$W^* = \arg\max_{W} P(W \mid O) = \arg\max_{W} P(O \mid W) P(W)$$

- Multiple-pass search
  - The 1st pass : the best syllable sequence & boundaries
  - The 2nd pass : multiple syllable candidates
  - The 3rd pass : word graph construction
  - The 4th pass : word (character) sequence





Mandarin Chinese
Spoken Document Retrieval (SDR)

From Speech Recognition to Spoken Document Retrieval

Task Definition of Spoken Document Retrieval

- Automatically ***indexing*** a collection of spoken documents with speech recognition techniques
- ***Retrieving*** relevant documents in response to a text/speech query

Why Is It An Important Problem

- Massive quantities of spoken audio are becoming available
- More people want to access and use this information
- Speech is currently a difficult media to browse and search
- There is a significant impact on the use of speech as a data type

Subword vs. Word for Retrieval

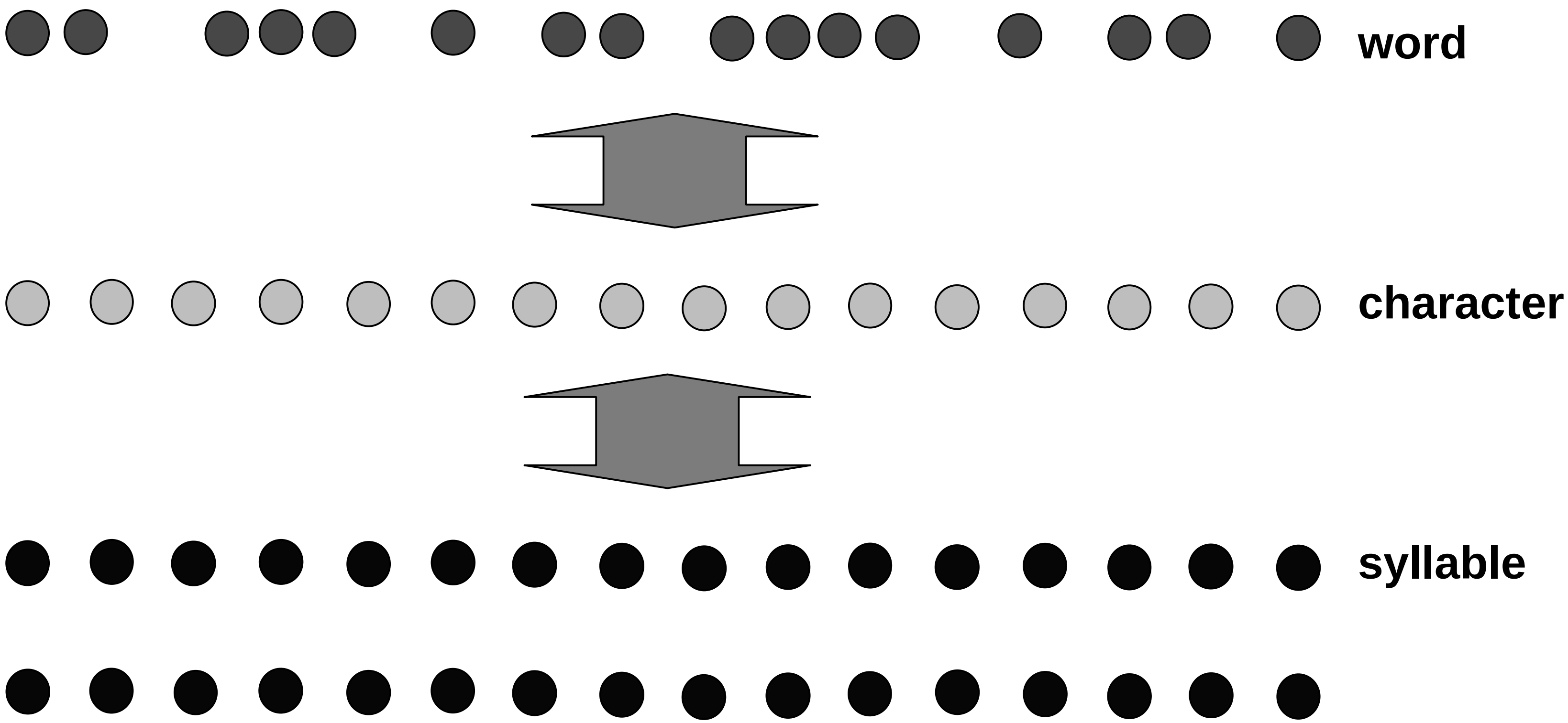
- ❑ Words contain lexical knowledge

Words enhance precision

- ❑ Subwords offer robustness against
  - ➢ Word tokenization ambiguity, e.g. 這一晚會如常舉行
    - – 這一晚 會 如常 舉行 [Tonight it will proceed as usual]
    - – 這一 晚會 如常 舉行 [This banquet will proceed as usual]
  - ➢ Open vocabulary problem
    - – An unlimited number of words, but 6,800 (or 13,000) characters and 400 syllables offer complete textual and phonological coverage for Mandarin Chinese
  - ➢ Homophone ambiguity
    - – 富庶 負數 複數 覆述 are totally different words but all pronounced as /fu shu/
    - – A foreign word may be translated into different Chinese words
  - ➢ Speech recognition errors

Subwords enhance recall

Multi-scale Indexing



Overlapping N-gram Indexing

Given a document (or a query)

u_1 u_2 u_3 u_4 u_5 u_6 u_7 u_8

$\bullet$ $\bullet$ $\bullet$ $\bullet$ $\bullet$ $\bullet$ $\bullet$ $\bullet$ . . . .

u_i can be a word or a subword (character, syllable, phoneme, etc.)

We can have

Uni-gram: $u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, \ldots$

Overlapping bi-gram: $u_1u_2, u_2u_3, u_3u_4, u_4u_5, u_5u_6, u_6u_7, u_7u_8, \ldots$

Overlapping tri-gram: $u_1u_2u_3, u_2u_3u_4, u_3u_4u_5, u_4u_5u_6, u_5u_6u_7, u_6u_7u_8, \ldots$

…..

Each overlapping N-gram is called an indexing term

Vector Space Model – Basic Idea

Document (query) represented by a feature vector

$$V = (ws_V(1), ws_V(2), \dots, ws_V(t), \dots, ws_V(T))$$

For a specific indexing term *t*, the weight is

$$ws_V(t) = (1 + \log(c_t)) \times \log(N_D / N_{D_t})$$

tf x idf

Query-document similarity is

$$S(q,d) = \frac{V_q \cdot V_d}{\|V_q\| \|V_d\|}$$

Vector Space Model – Information Fusion

$$WS_{V_{qi}}(t) = (1 + \log(\sum_{j=1}^{n_t} cm_t(j)))$$

Using confidence measures instead of frequency counts

$$WS_{V_{qi}}(t) = (1 + \log(\sum_{j=1}^{n_t} cm_t(j))) \cdot \log(N_D / N_{D_t})$$

$$S_i(q,d) = \frac{V_{qi} \cdot V_{di}}{\left\|V_{qi}\right\| \left\|V_{di}\right\|}$$

$$S(q,d) = \sum_{i=1}^{I} w_i S_i(q,d)$$

Probability Model

- Given a user-generated query and a set of documents, we wish to rank the documents according to the probability that *D* is relevant, conditioned on the fact that the user produced *Q*; i.e., $P(D \text{ is } R \mid Q)$

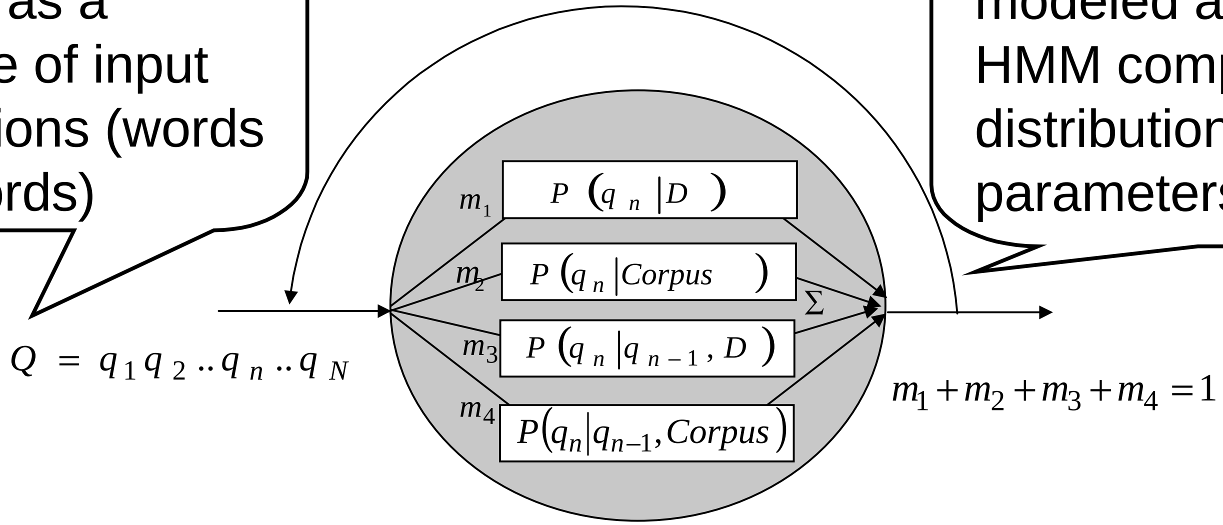
$$D^* = \underset{D}{\text{argmax}}\ P(D \text{ is } Relevant \mid Q) = \underset{D}{\text{argmax}}\ \frac{P(Q \mid D \text{ is } Relevant)P(D \text{ is } Relevant)}{P(Q)}$$

$$= \underset{D}{\text{argmax}}\ P(Q \mid D \text{ is } Relevant)P(D \text{ is } Relevant) \cong \underset{D}{\text{argmax}}\ P(Q \mid D \text{ is } Relevant)$$

HMM/N-Gram-Based Retrieval Model

Each query is modeled as a sequence of input observations (words or subwords)

Each document is modeled as a discrete HMM composed of distributions of n-gram parameters



$$P(Q|D \text{ is } R) = \left[m_1 P(q_1|D) + m_2 P(q_1|Corpus)\right] \cdot \prod_{n=2}^{N}\left[m_1 P(q_n|D) + m_2 P(q_n|Corpus) + m_3 P(q_n|q_{n-1}, D) + m_4 P(q_n|q_{n-1}, Corpus)\right]$$

HMM/N-Gram-Based Retrieval Model (cont'd)

- Simplified to Unigram-Based only

$$P(Q|D \text{ is } R) = \prod_{n=1}^{N} \left[m_1 P(q_n|D) + m_2 P(q_n|Corpus) \right]$$

- Extended to Unigram-/Bigram-/Trigram-Based

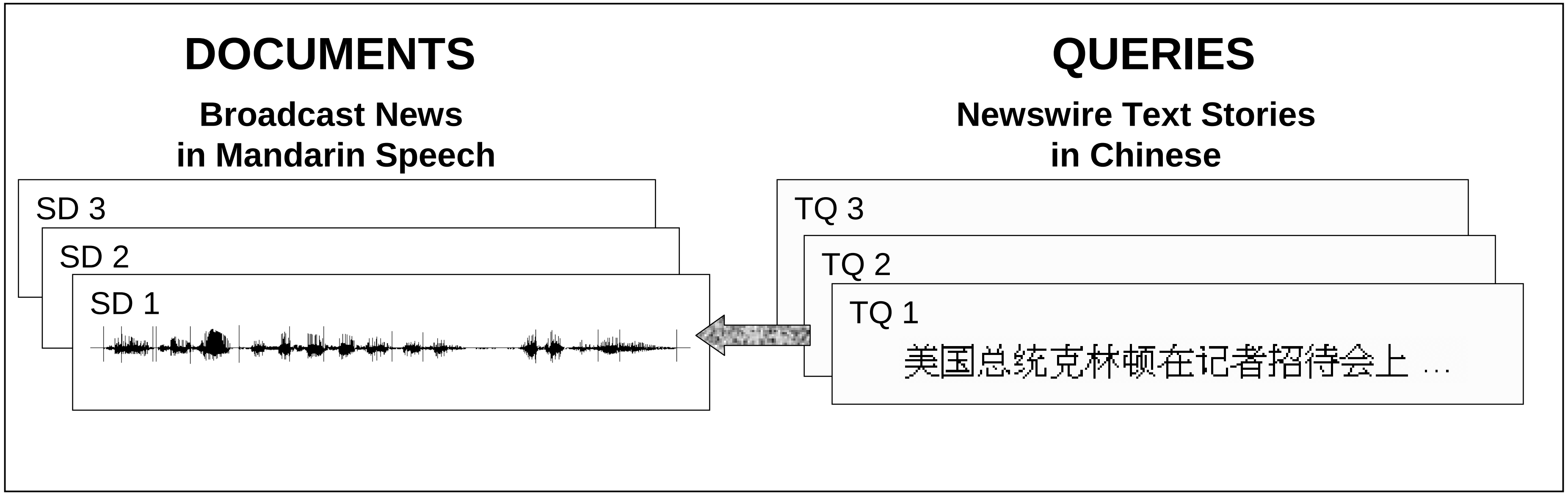
$$\begin{aligned} P(Q|D \text{ is } R) = & \left[m_1 P(q_1|D) + m_2 P(q_1|Corpus) \right] \\ & \cdot \left[m_1 P(q_2|D) + m_2 P(q_2|Corpus) + m_3 P(q_2|q_1, D) + m_4 P(q_2|q_1, Corpus) \right] \\ & \cdot \prod_{n=3}^{N} \left[m_1 P(q_n|D) + m_2 P(q_n|Corpus) + m_3 P(q_n|q_{n-1}, D) + m_4 P(q_n|q_{n-1}, Corpus) \right. \\ & \left. + m_5 P(q_n|q_{n-2}, q_{n-1}, D) + m_6 P(q_n|q_{n-2}, q_{n-1}, Corpus) \right] \end{aligned}$$

$P(q_n|Corpus), P(q_n|q_{n-1}, Corpus)$ N-gram probabilities estimated from a corpus for modeling the general distribution of the indexing terms

m_i can be estimated using the expectation-maximization (EM) algorithm, and all the documents share the same weights

Retrieval Context

- An entire newswire text story in Chinese as a query and broadcast news in Mandarin speech as documents - *Query-By-Example*



Experimental Corpora (I)

- Topic Detection and Tracking corpora (TDT-2 & TDT-3) from Linguistic Data Consortium (LDC)
  - TDT-2 as the development test set while TDT-3 as the evaluation test set
  - Spoken documents: broadcast news in Mandarin from Voice of American (VOA)
  - Text queries: text news stories in Chinese from Xinhua News Agency

	TDT-2 (Development) 1998, 02~06			TDT-3 (Evaluation) 1998, 10~12		
# Spoken documents	2,265 stories, 46.03 hrs of audio			3,371 stories, 98.43 hrs of audio		
# Distinct text queries (query-by-example)	16 Xinhua text stories (Topics 20001~20096)			47 Xinhua text stories (Topics 30001~30060)		
	Min.	Max.	Mean	Min.	Max.	Mean
Document length (characters)	23	4841	287.1	19	3667	415.1
Query length (characters)	183	2623	532.9	98	1477	443.6
Number of relevant documents/query	2	95	29.3	3	89	20.1

Experimental Corpora (II)

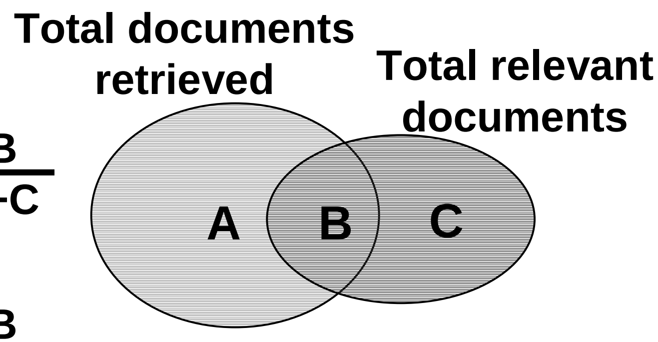
- An outside text corpus consisting of 40 million Chinese characters for estimating the corpus N-gram probabilities
  $P(q_n|Corpus), P(q_n|q_{n-1}, Corpus)$
- An outside training query set consisting of 819 query exemplars and their corresponding query-document relevance information with respect to the development set of the TDT-2 document collection for training the weights m_i
- A pronunciation lexicon (~50k words)
  - LDC Mandarin Chinese Lexicon + 24k words extracted from Dragon's word recognition output
- Speech recognition error rates (Dragon's recognizer)
  - TDT-2: 35.38% (word), 17.69% (character), 13.00% (syllable)
  - TDT-3: 36.97% (word), 19.78% (character), 15.06% (syllable)

IR Performance Measures

- Recall and Precision

$$\text{Recall} = \frac{\text{relevant documents retrieved}}{\text{total relevant documents}} \quad \frac{B}{B+C}$$

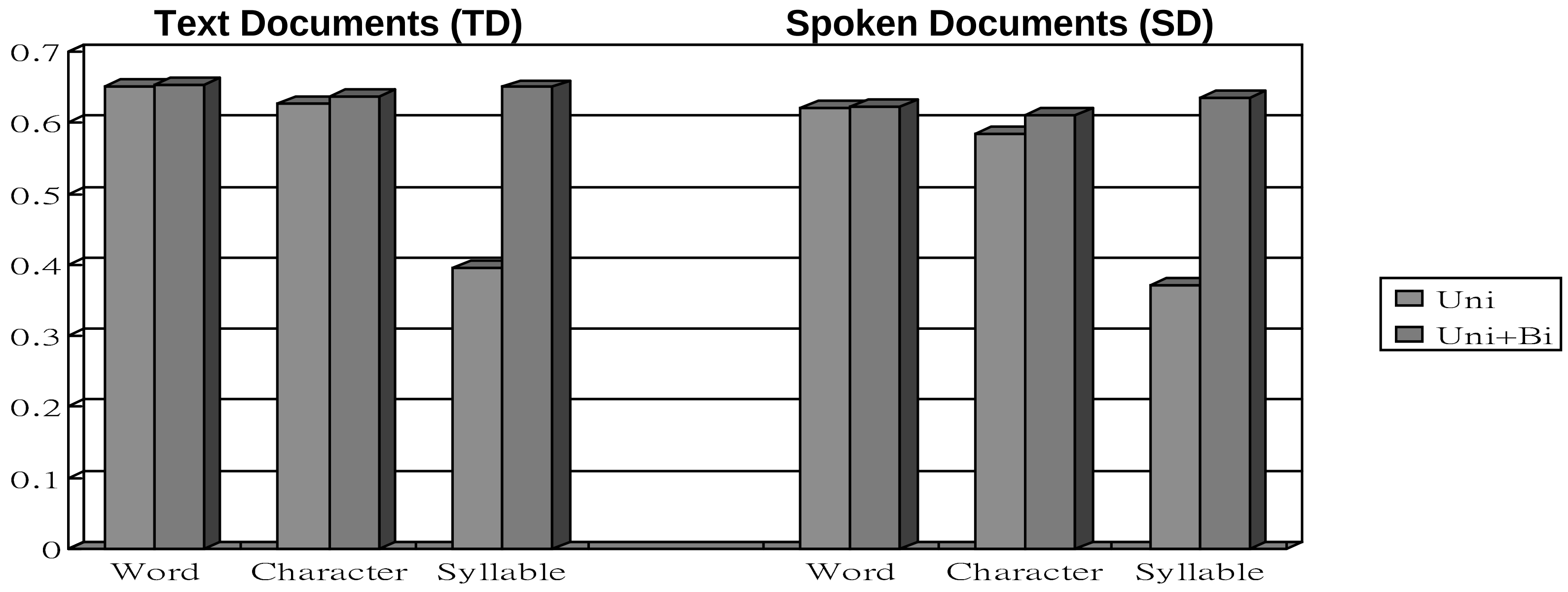
$$\text{Precision} = \frac{\text{relevant documents retrieved}}{\text{total documents retrieved}} \quad \frac{B}{A+B}$$



- Mean Average Precision (mAP)
  - Precision averaged at relevant documents and across queries
    - e.g. relevant documents ranked at 1, 5, 10, precisions are 1/1, 2/5, 3/10, non-interpolated average precision=(1/1+2/5+3/10)/3
    - mAP= $\frac{1}{|Q|}\sum_{q=1}^{|Q|}(\text{non-interpolated average precision})_q$

Experimental Results (I)

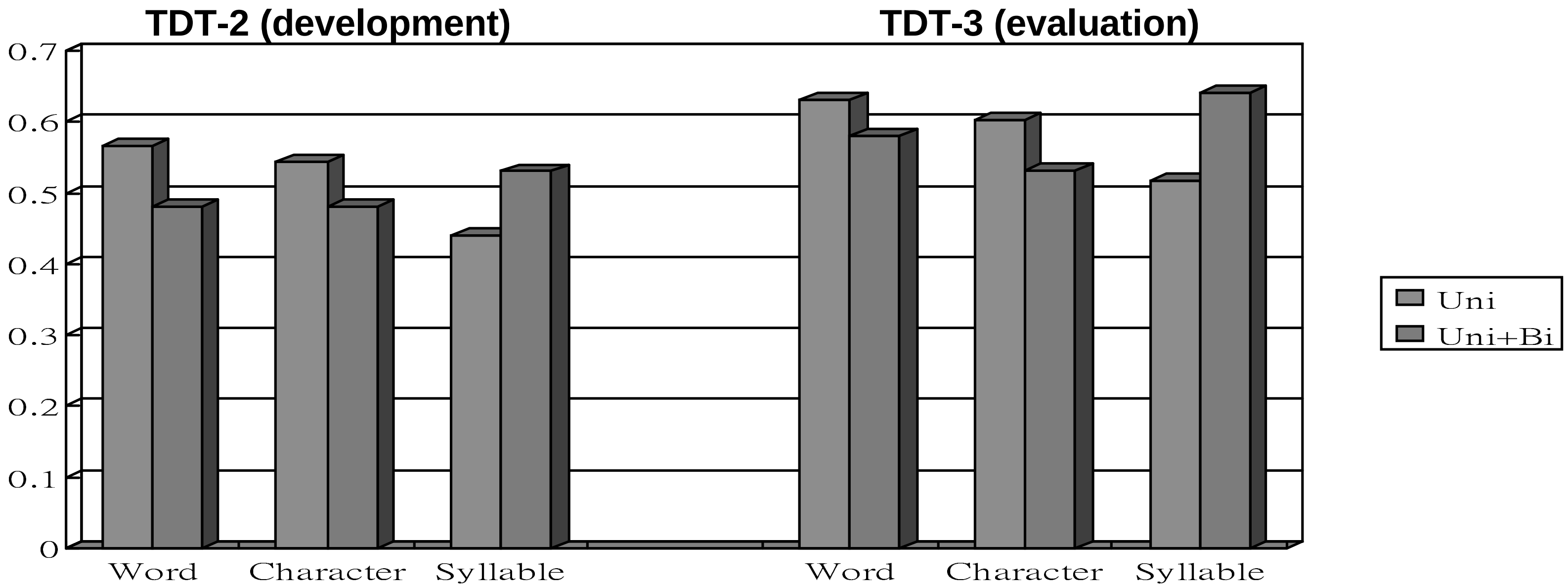
- Vector-Space-Based Retrieval Model on the evaluation set



1. Subword indexing features performed as well as word indexing features
2. Bigram information did help, in particular in the syllable case
3. The SD cases were only slightly worse than the TD cases (wer>35%)

Experimental Results (II)

- HMM/N-Gram-Based Retrieval Model

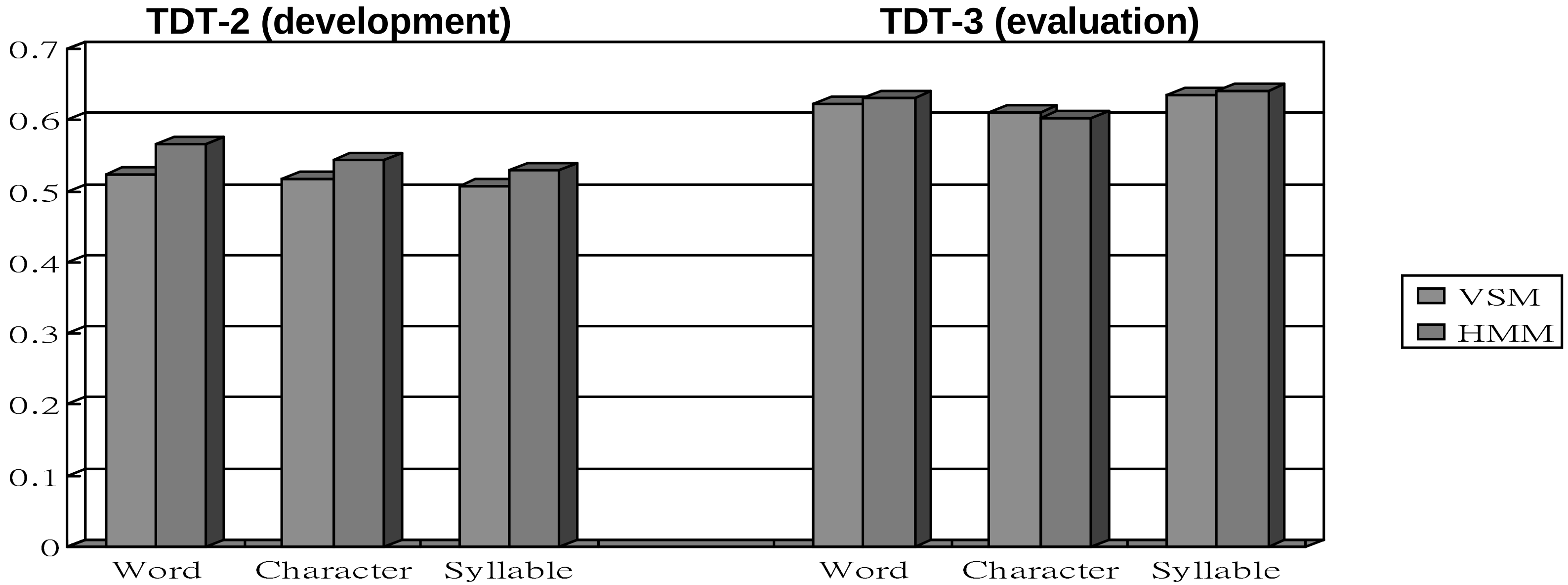


1. Word indexing features outperformed subword indexing features in most cases, but syllable indexing features performed very well in TDT-3

2. Bigram information did not help in word and character cases

Experimental Results (III)

- Comparison of two models **VSM:Uni+Bi** **HMM:Uni(Word,Character) Uni+Bi(Syllable)**

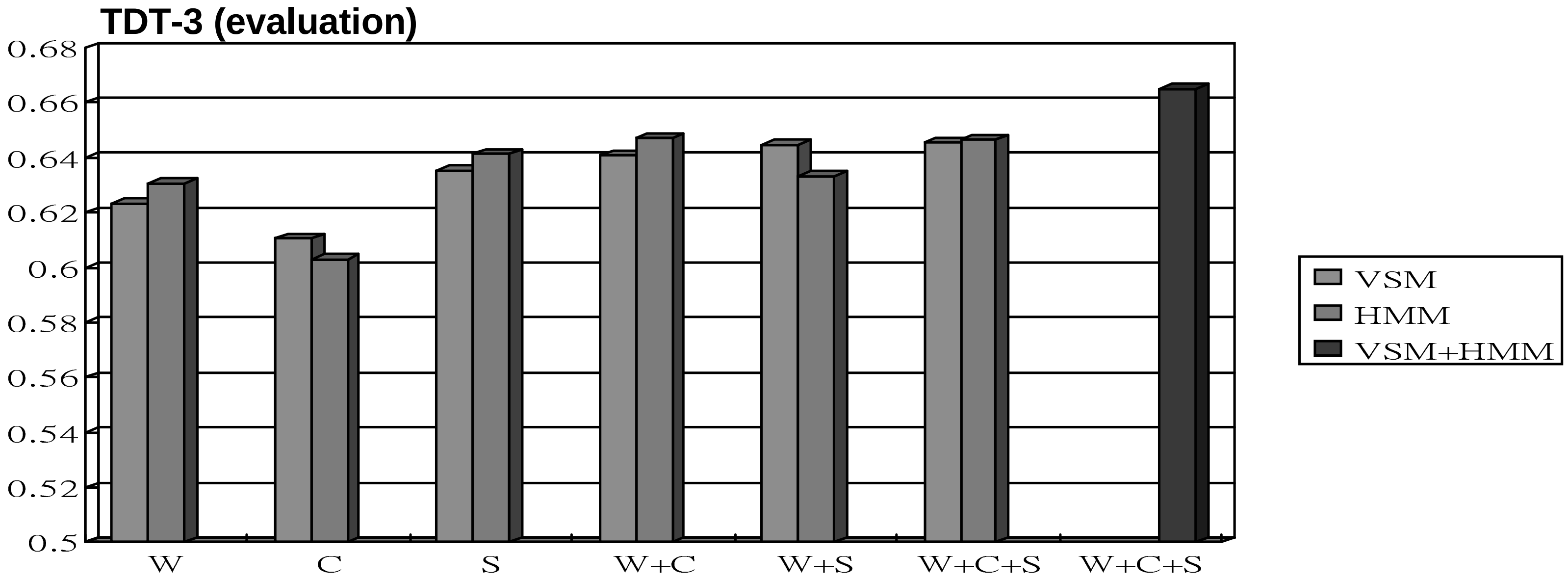


1. The HMM/N-gram-based approach achieved consistently better performance than the vector space model approach
2. The difference between the two was larger for the TDT-2 development set from which the weights were trained

Experimental Results (IV)

- Information fusion

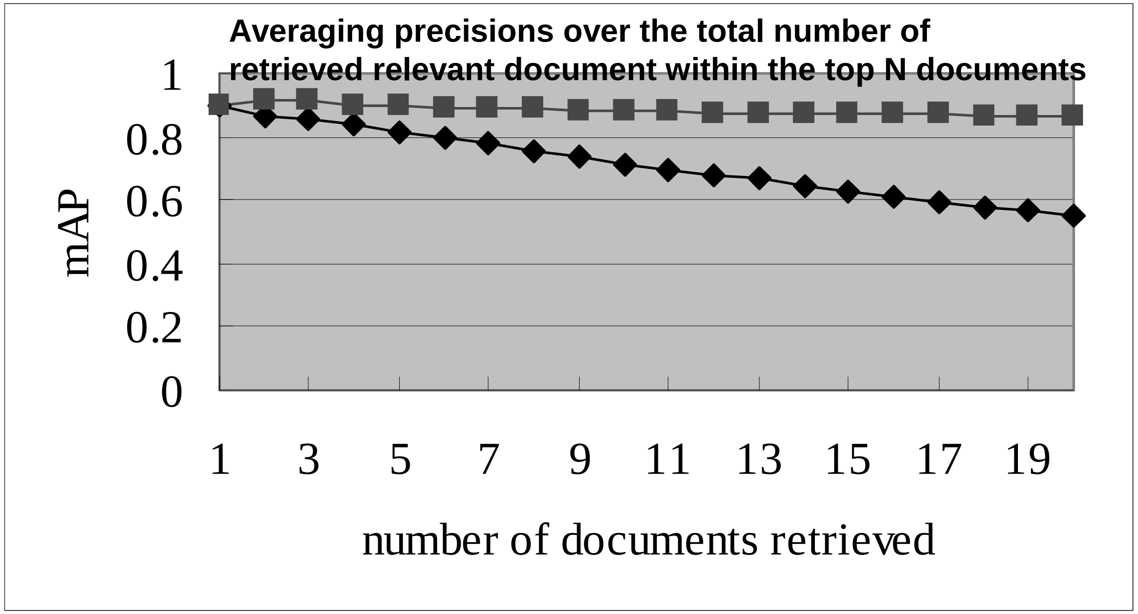
VSM:Uni+Bi HMM:Uni(Word,Character)
Uni+Bi(Syllable)



1. Fusion of different indexing features was in general helpful for retrieval
2. Fusion of different model approaches was helpful as well

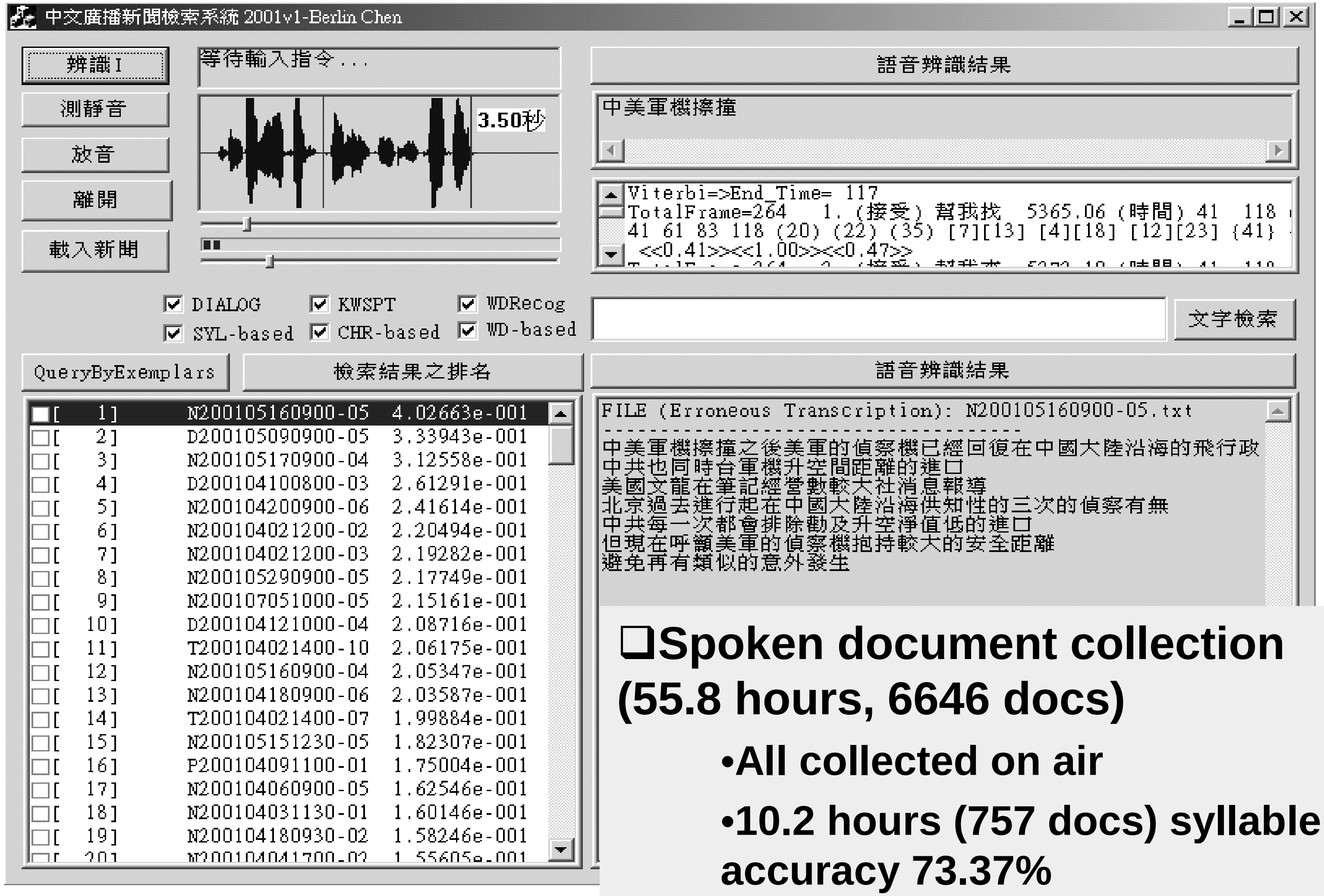
SoVideo – Mandarin Chinese SDR

SoVideo – Performance Evaluation



- ❑Spoken document collection (53.3 hours, 2617 docs)
  - •10.2 hours collected on air (syllable accuracy 73.37%)
  - •43.1 hours in RealAudio format (syllable accuracy 27.87%)
- ❑Text queries only (40 short queries, 2-7characters/query)
  - •90% of queries get the relevant document with only one returned
  - •95% of queries get at least 1 relevant within 3

Prototype System



- **Spoken document collection (55.8 hours, 6646 docs)**
  - **All collected on air**
  - **10.2 hours (757 docs) syllable accuracy 73.37%**

Conclusions

- We have investigated the use of words, characters and syllables in audio indexing for Mandarin Chinese spoken document retrieval
  - Word-level indexing features outperformed character- and syllable-level features in most cases
  - Syllable-level indexing features performed very well in the real, desired case of retrieval from the erroneous speech transcriptions (SD) of the evaluation set
- The HMM/N-gram-based retrieval model is in general better than the Vector-space-based retrieval model
- Fusion of indexing features of different levels is in general helpful for retrieval
- Fusion of different model approaches is helpful as well

Thank You!