

美

**Mandarin-English Information (MEI):
Investigating Translingual Speech Retrieval**

**Johns Hopkins University
Center of Language and Speech Processing
Summer Workshop 2000
Progress Update**

The MEI Team
August 2, 2000

Outline

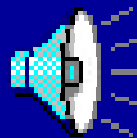
- Baseline (Pat, Gina, Wai-Kit)
- Upper Bounds (Pat, Erika, Helen)
- Climbing Upwards (Upcoming Research Problems)
 - translation (Gina, Jian Qiang)
 - word-subword fusion (Helen, Doug, Wai-Kit)
 - named entities , numerals (Helen, Sanjeev, Wai-Kit, Karen)
 - syllable lattice generation (Hsin-Min, Berlin)

The MEI Task

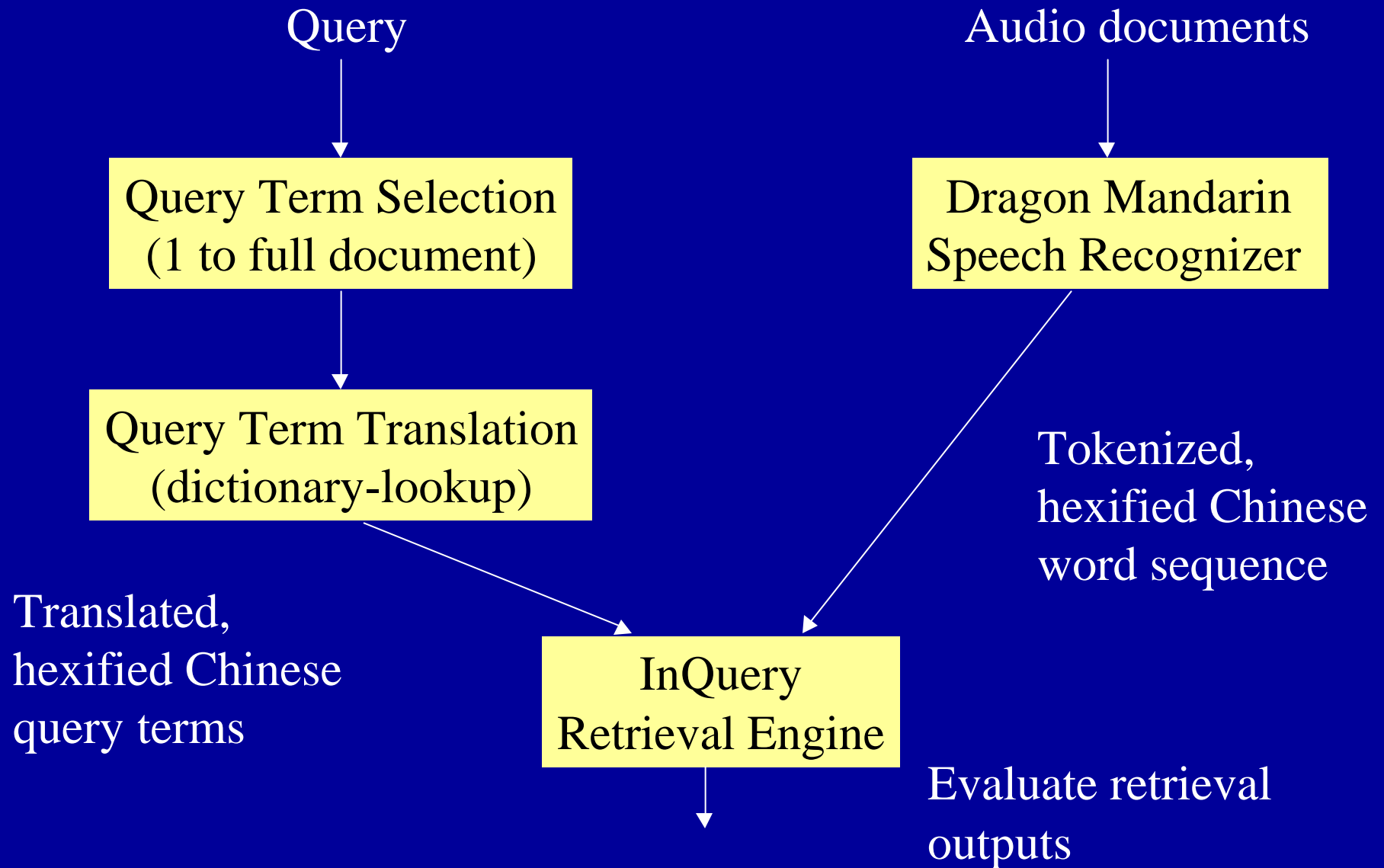
- An example query (NYT, AP newswire)

A China Airlines A-310 jetliner returning from the Indonesian island of Bali with 197 passengers and crew crashed and burst into flames Monday night just short of Taipei's Chiang Kai-Shek Airport.....
(full story used as query, typically 200-500 words)

- An example document (VOA)
 - accompanied by raw anchor scripts



Our Baseline System



Our First Retrieval Experiment...

- Queries
 - 17 exemplars
 - 1 per topic in TDT2 corpus
- Documents
 - 2265 in all
 - ~500 belong to at least 1 topic
 - others are “off-topic” or “briefs”
 - each topic has ≥ 2 relevant documents

Our First Retrieval Experiment

- No. of query terms selected = 100 (sweep)
- No. of alternative translations per term = 1
- Word-based retrieval
- Average Precision = 16.91%

In Search of Upper Bounds...

- Confounding factors on query side
 - term selection
 - translation (no. of terms, definition of a term, named entities, dictionary / COTS system)
- Confounding factors on the document side
 - syllable recognition performance, OOV
 - word tokenization
- Confounding factors in retrieval
 - word-based or subword-based (characters, syllables)
 - subword n-grams (n=??)

Upper Bounds (Word)

- Queries (ASR); Documents (ASR)
 - isolates the confounding factors (term selection, translation, recognition performance, word tokenization)
 - Ave Precision=73.3%
- Queries (Xinhua); Documents (ASR/TKN)
 - isolate similar confounding factors
 - resembles MEI TDT task (queries and documents come from different news sources)
 - word tokenization (CETA / Dragon)
 - Best Ave Precision = 53.5%(ASR), 58.7% (TKN)

Chinese Words and Subwords

- Characters (written) -> syllables (spoken)

- Degenerate mapping

- 行 /hang2/, /hang4/, /heng2/ or /xing2/

- /fu4 shu4/ (LDC's CALLHOME lexicon)

富庶 負數 復數 覆述

- Tokenization / Segmentation

- /zhe4 yi1 wan3 hui4 ru2 chang2 ju3 xing2/

這一晚 會 如常 舉行

這一 晚會 如常 舉行

這一 晚會 如 常 舉行

Upper Bounds (Subword)

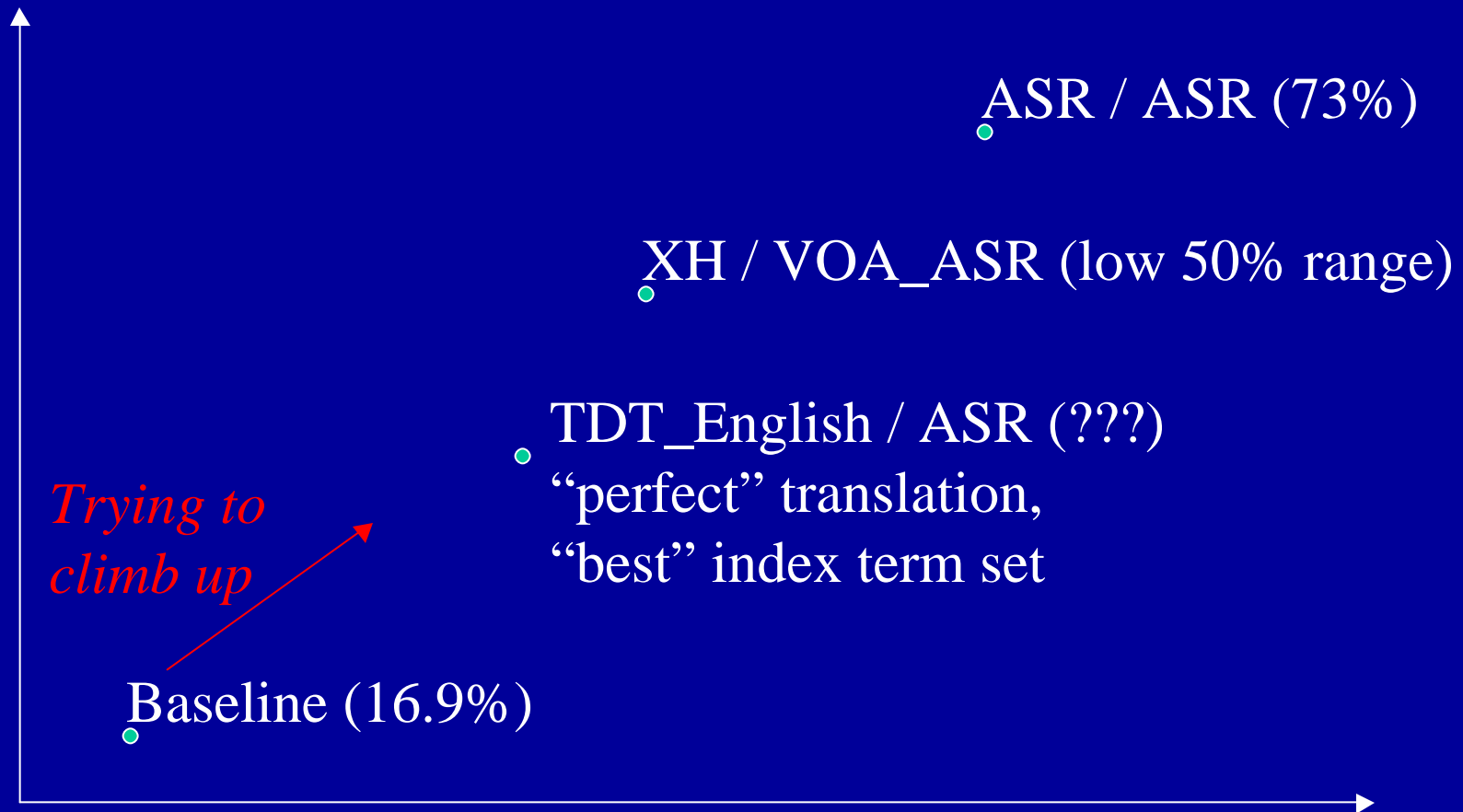
- Queries (Xinhua); Documents (ASR/TKN)
 - character-based retrieval
 - overlapping character n-grams (document, within-term for queries, bigrams fare best)
 - Best Ave Precision = 54.3%(ASR), 55.9%(TKN)
 - overlapping bigrams in queries
 - Best Ave Precision = 61.7% (cross-term overlap)
 - syllable-based retrieval
 - word tokenization affects syllable lookup
 - syllable bigrams fare best
 - Best Ave Precision = 51.6%(ASR), 53.3% (TKN)

Upper Bound (Translingual)

- Putting back the **translingual** element
- Selected English query terms --> translated Chinese query terms (Oracle -- Jian Qiang Wang)
- Retrieval performance
 - word-based (180 terms, no #syn, #sum) **50.6%**
 - subword-based retrieval (character bigrams, #sum **52.1%**, #syn **52.3%**)
 - TKN??

Thus Far...

Ave Precision



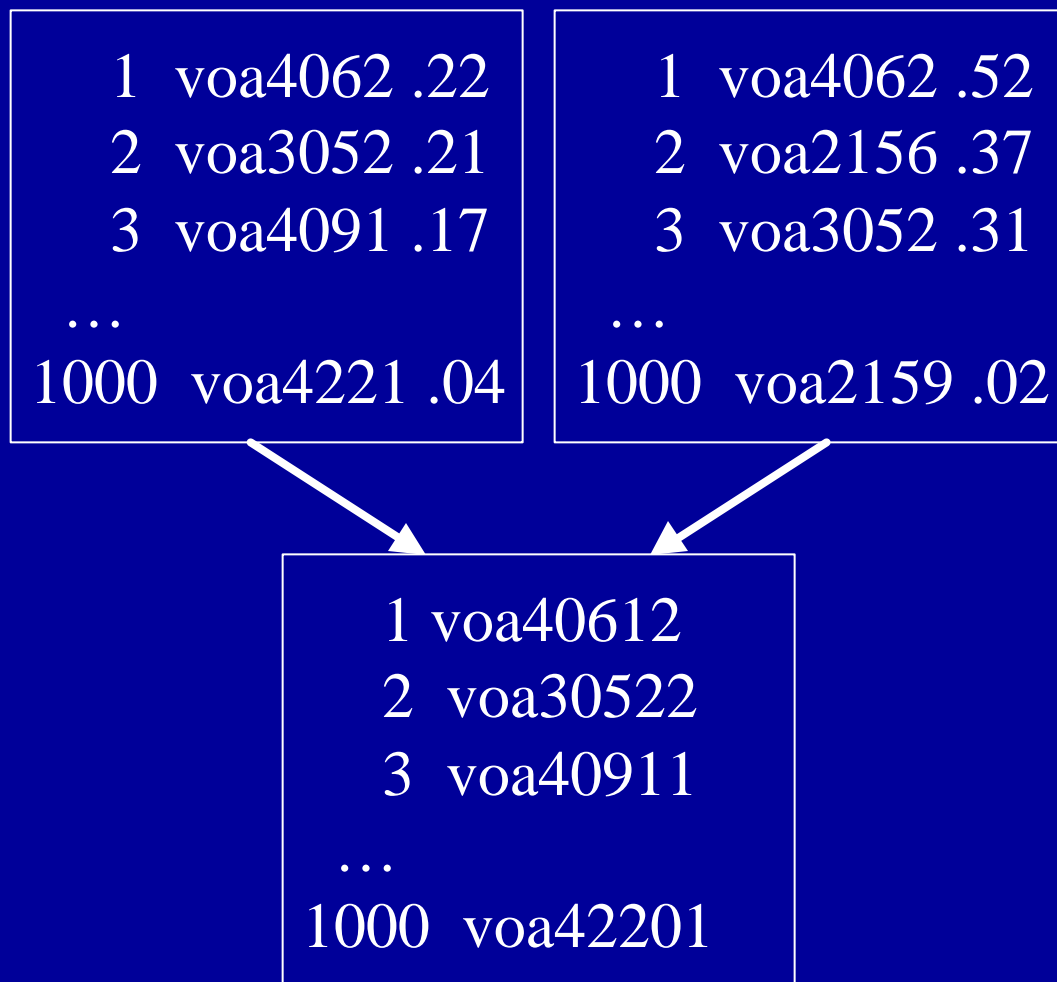
Better Translation

- # translation alternatives per term
 - Current best (120 query terms, 3 translations per term, word-based retrieval, ASR reseg with CETA, #sum 28.1%)
 - (90 query terms, 2 translations pre term, word-based retrieval, ASR orig #sum 27.53%)
- Phrase-based translation
 - 2 types of phrases (named entities, dictionary-based phrases)
 - term selection (consider *both* phrases and component words), higher # terms
 - Current best (250 query terms, all translations, word-based retrieval, 43.3%)

Word-Subword Fusion

- Words incorporate lexical knowledge
- Subwords are intended to handle the OOV problem
- Combination of both may beat either alone
- Ranked list of retrieved documents
 - from word-based retrieval
 - from subword-based retrieval

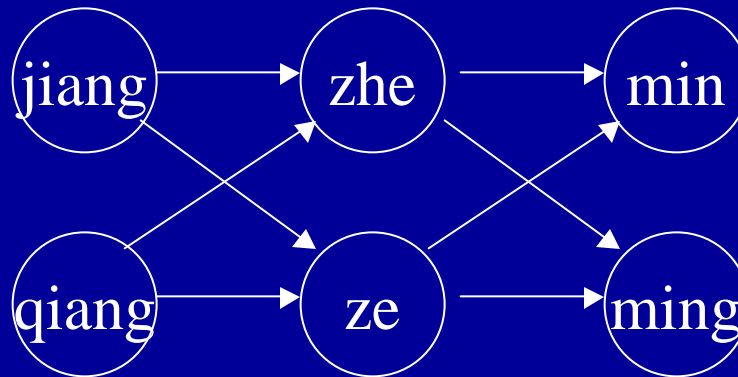
Merging: Loose Coupling



- Types of Evidence
 - Score
 - Rank
- Score Combination
 - Max
 - Linear combination
- Rank Combination
 - Round robin
 - Source bias
 - Query bias

Tight Coupling: Words and Bigrams

Lattice:



Words: Jiang Zemin

Words: Jiang Zemin

Bigrams: jiang_zhe jiang_ze qiang_zhe qiang_ze
zhe_min zhe_ming ze_min ze_ming

Combination: jiang_zhe zhe_min Jiang Zemin

Word-Subword Fusion (weighted similarity)

- Merging ranked lists
- Each retrieved document is scored

$$S(Q, D) = \sum_i w_i S_i(Q_i, D_i)$$

– i denotes words, subword n-grams

Numerals and Named Entities

- Verbalize numerals
- Named Entities
 - BBN tags (names of locations, people, organization)
 - Derive Bilingual Term List from TDT2
 - English letter-to-phone generation
 - Cross-lingual phonetic mapping (English phones to Chinese phones)
 - Syllabification

Cross-Lingual Phonetic Mapping

Named entity Jiang Zemin, Kosovo

Syllabify Pinyin Spelling

E.g. jiang ze min

English Pronunciation Lookup
or
Letter-to-Phone Generation

English Phones, e.g. k a o s a x v o w

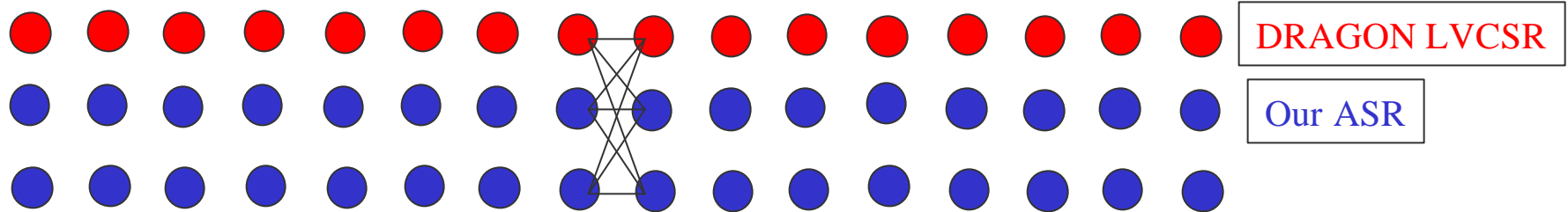
Cross-lingual Phonetic Mapping

Chinese Phones, e.g. k e s u o w o

Syllabification

Chinese syllables, e.g. ke suo wo

Syllable Lattice for Document Representation



- Address ASR errors and OOV
 - Augment Dragon ASR output with alternate syllable hypotheses
- Generate syllable n-grams for audio indexing
- Include into word-subword fusion

Lots to do still...

