

美

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

**Johns Hopkins University
Center of Language and Speech Processing
Summer Workshop 2000
Opening Ceremony**

The MEI Team
July 17, 2000

MEI Team

- Senior Members

Helen Meng	Chinese University of Hong Kong
Erika Grams	Advanced Analytic Tools
Sanjeev Khudanpur	Johns Hopkins University
Gina-Anne Levow	University of Maryland
Douglas Oard	University of Maryland
Patrick Schone	US Department of Defense
Hsin-Min Wang	Academia Sinica, Taiwan

- Students

Berlin Chen	National Taiwan University
Wai-Kit Lo	Chinese University of Hong Kong
Karen Tang	Princeton University
Jianqiang Wang	University of Maryland

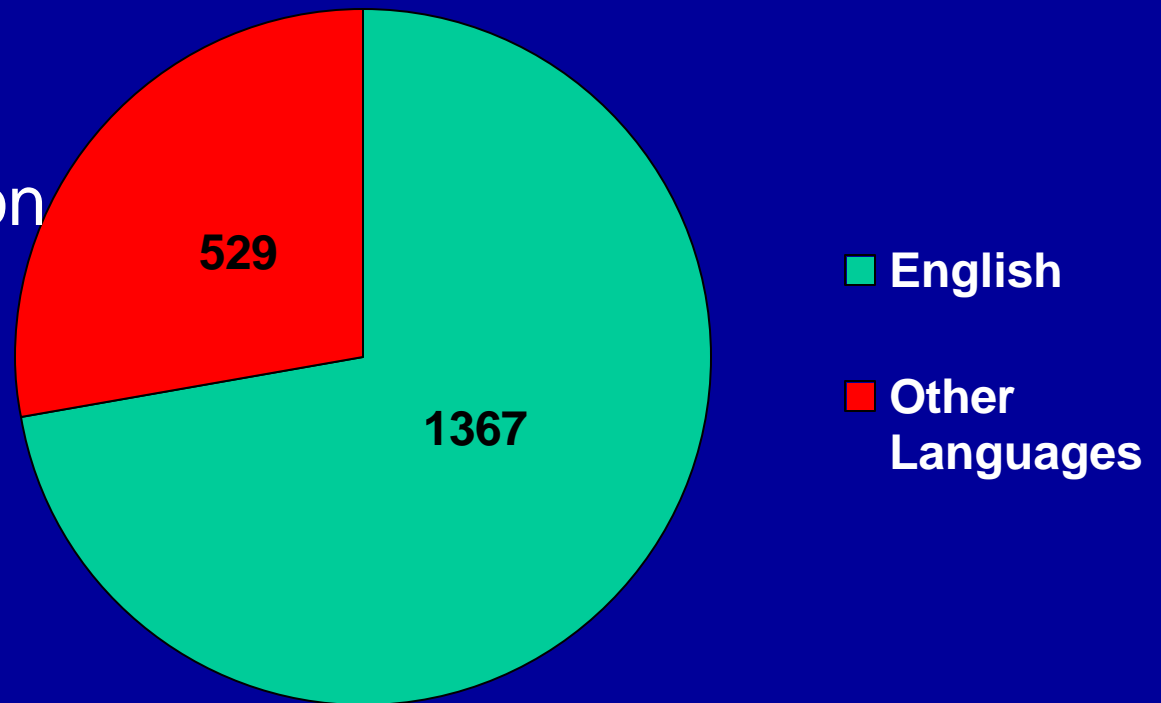
Outline

- Motivation
- MEI project overview
- Research challenges
- System architecture

Motivation

- Monolingual speech retrieval applications are emerging, e.g.
 - <http://speechbot.research.compaq.com>

Internet-accessible
Radio and Television
Stations

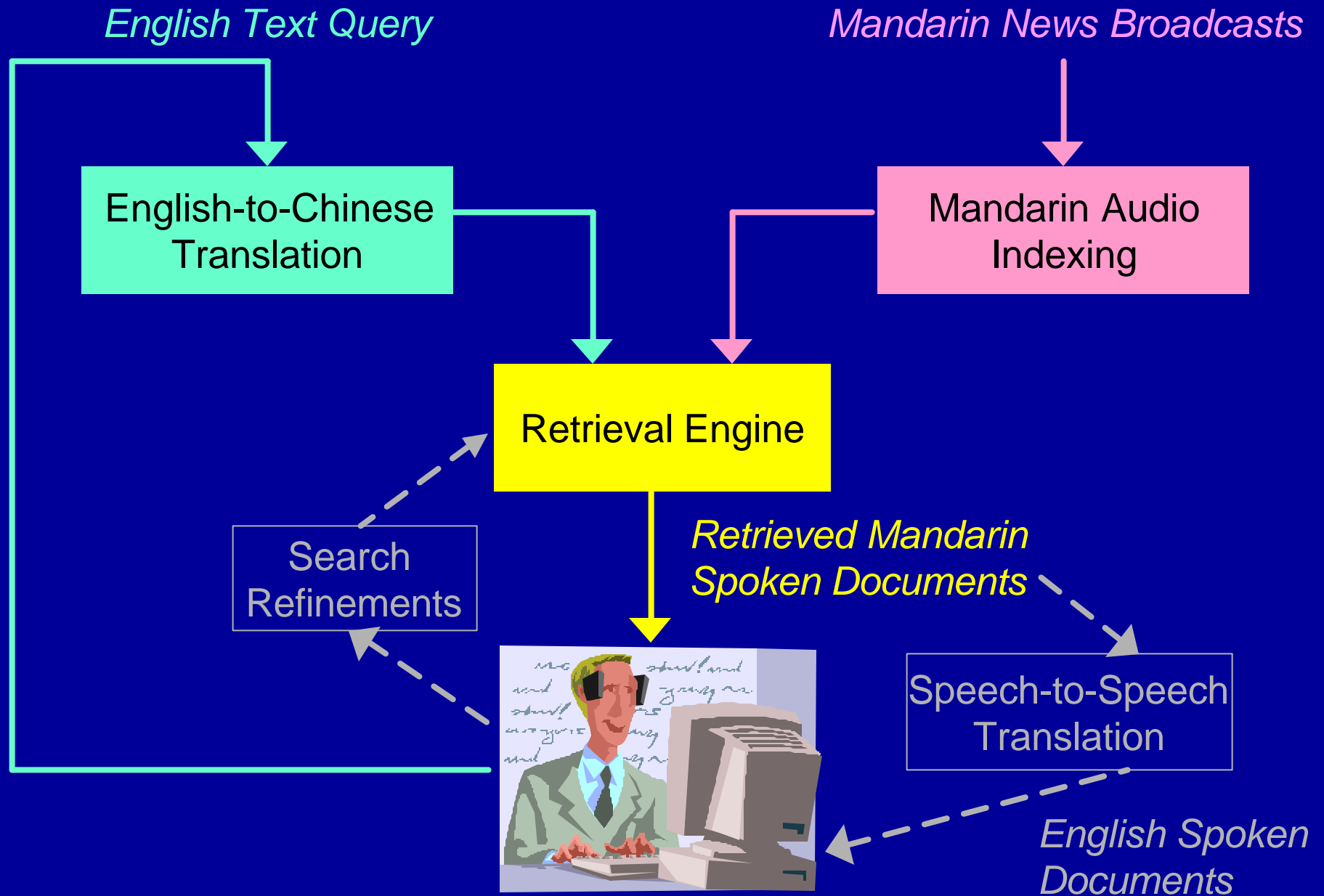


source: www.real.com, Feb 2000

Translingual Speech Retrieval

- Allow **anyone** to find information that is expressed in **any spoken language**

The Big Picture



Related Work

- TREC Spoken Document Retrieval
 - Close coupling of recognition and retrieval
- TREC Cross-Language Retrieval
 - Close coupling of translation and retrieval
- TDT-3 Topic Tracking
 - Coupling recognition, translation and retrieval
 - Using speech recognition transcripts

The MEI Project

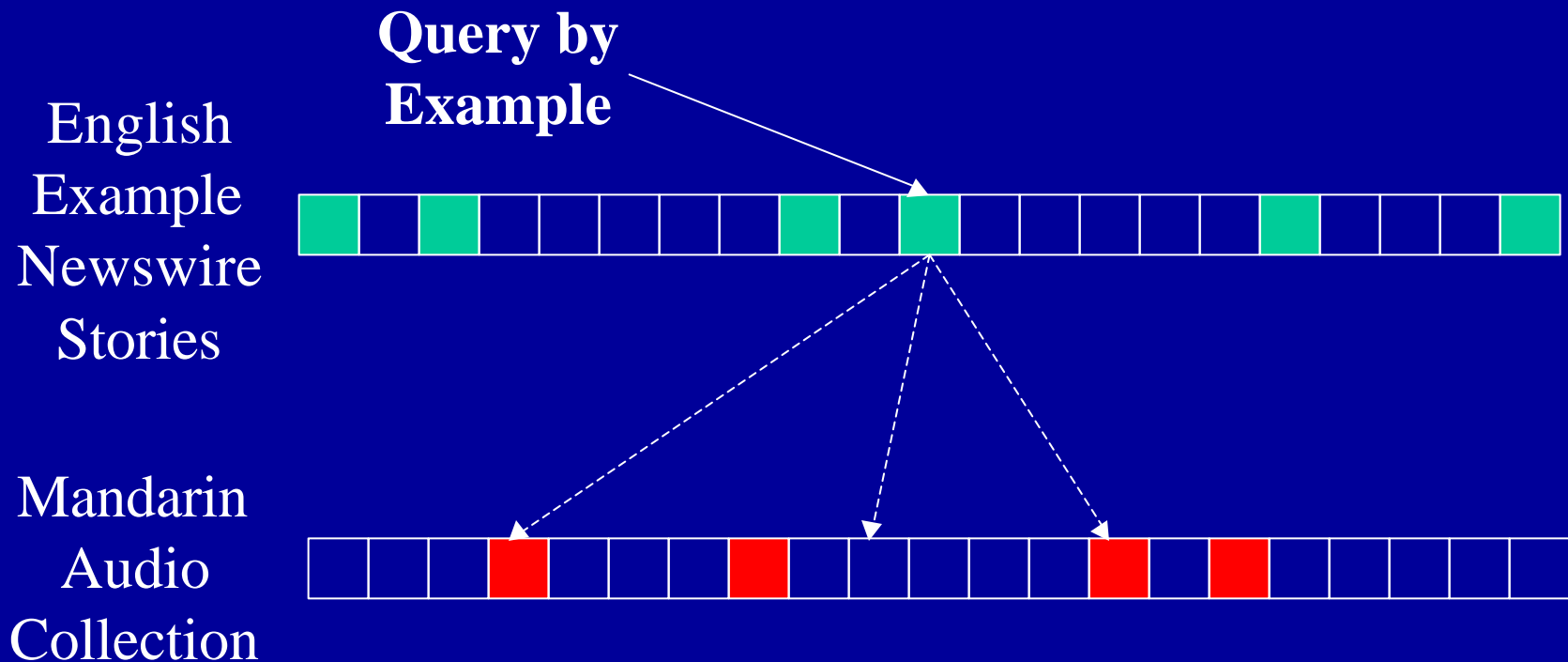
- Closely couple recognition and translation
 - for the purpose of retrieval

Research Challenges

- Multi-scale audio indexing
 - Multiple feature sets capture more information
- Multi-scale translation
 - Lexicon and pronunciation are complementary
- Multi-scale retrieval
 - Combination of evidence can add robustness

Task and Corpus

- TDT Corpus
- Challenge: using an English story as the query, find related Mandarin audio documents



Multi-scale Mandarin Processing: Phonological Considerations

- Chinese is a syllable-based language
- Mandarin is the official Chinese dialect
 - ~400 base syllables, 4 lexical tones + light tone
- Syllable structure **(CG)V(X)**
 - **(CG)**: onset, optional, consonant+medial glide
 - **V**: nuclear vowel
 - **X**: coda, glide / alveolar nasal / velar nasal
 - ~ 21 initials, 39 finals
- Circumvent the OOV problem

Multi-scale Mandarin Processing: Linguistic Considerations

- 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

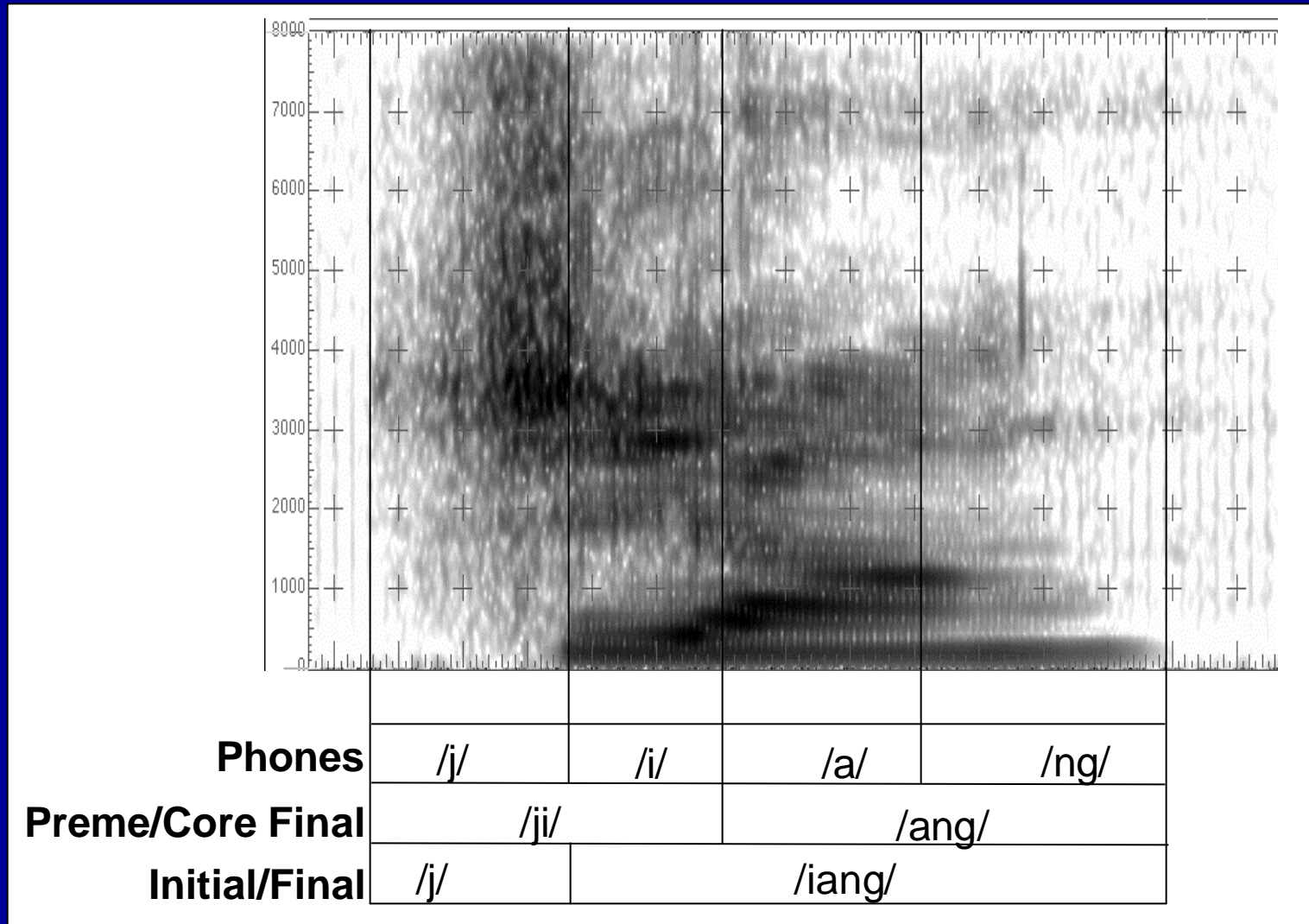
- Tokenization / Segmentation
 - /zhe4 yi1 wan3 hui4 ru2 chang2 ju3 xing2/

這一晚 會 如常 舉行

這一 晚會 如常 舉行

這一 晚會 如 常 舉行

Multi-scale Audio Indexing



Preme/Toneme, Initial/Tonal-Final

syllable /jiang/, single-character word, part of a multi-char word

Previous Work on Audio Indexing

- Syllable lattice matching [Chen, Wang & Lee, 2000]
- Comparison of multi-scale acoustic models [Lo, Meng & Ching, 2000]
- Overlapping syllable n -grams [Meng et al., 1999]
- Skipped syllable pairs [Chen, Wang & Lee, 2000]

Multi-scale Translation

- Word-scale
 - Phrase-based, e.g. “the White House”
 - Dictionary-based [Levow & Oard 00]
 - Lexical gaps: parallel corpora, comparable corpora
- Subword-scale
 - Named entities: transliteration by cross-lingual phonetic mapping
 - Northern Ireland 北愛爾蘭 /bei2 ai4 er3 lan2/
 - Kosovo (/ke1-suo3-wo4/, /ke1-suo3-fo2/, /ke1-suo3-fu1/, /ke1-suo3-fu2/)

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

Multi-scale Retrieval

- Word-scale exploits lexical knowledge
 - Enhances precision
- Subwords can achieve complete coverage
 - Enhances recall
- Combination of evidence may be best
 - If a good merging strategy can be found

Merging Strategies

- Loose coupling
 - Separate retrieval runs
 - Merge ranked lists [Voorhees 1995]
- Tight coupling
 - Unified indexing of words and subwords
 - Single ranked list
 - [Ng 2000]

Multi-scale Retrieval Techniques

- Subword-scale
 - Syllable lattice matching [Chen, Wang & Lee 2000]
 - Overlapping syllable n -grams [Meng et al. 1999]
 - Syllable confusion matrix [Meng et al. 1999]
- Word-scale
 - Structured queries [Pirkola 1998]
 - Structured translation [Sperer & Oard 2000]
- Robust Retrieval
 - speech recognition errors
 - translation / transliteration ambiguities

Research Plan

- *“MEI: Mandarin-English Information”*,
Proceedings of the TDT Workshop, 2000.
- *“Mandarin-English Information (MEI):
Investigating Translingual Retrieval”*
Proceedings of the NAACL Workshop on
Embedded Machine Translation, 2000



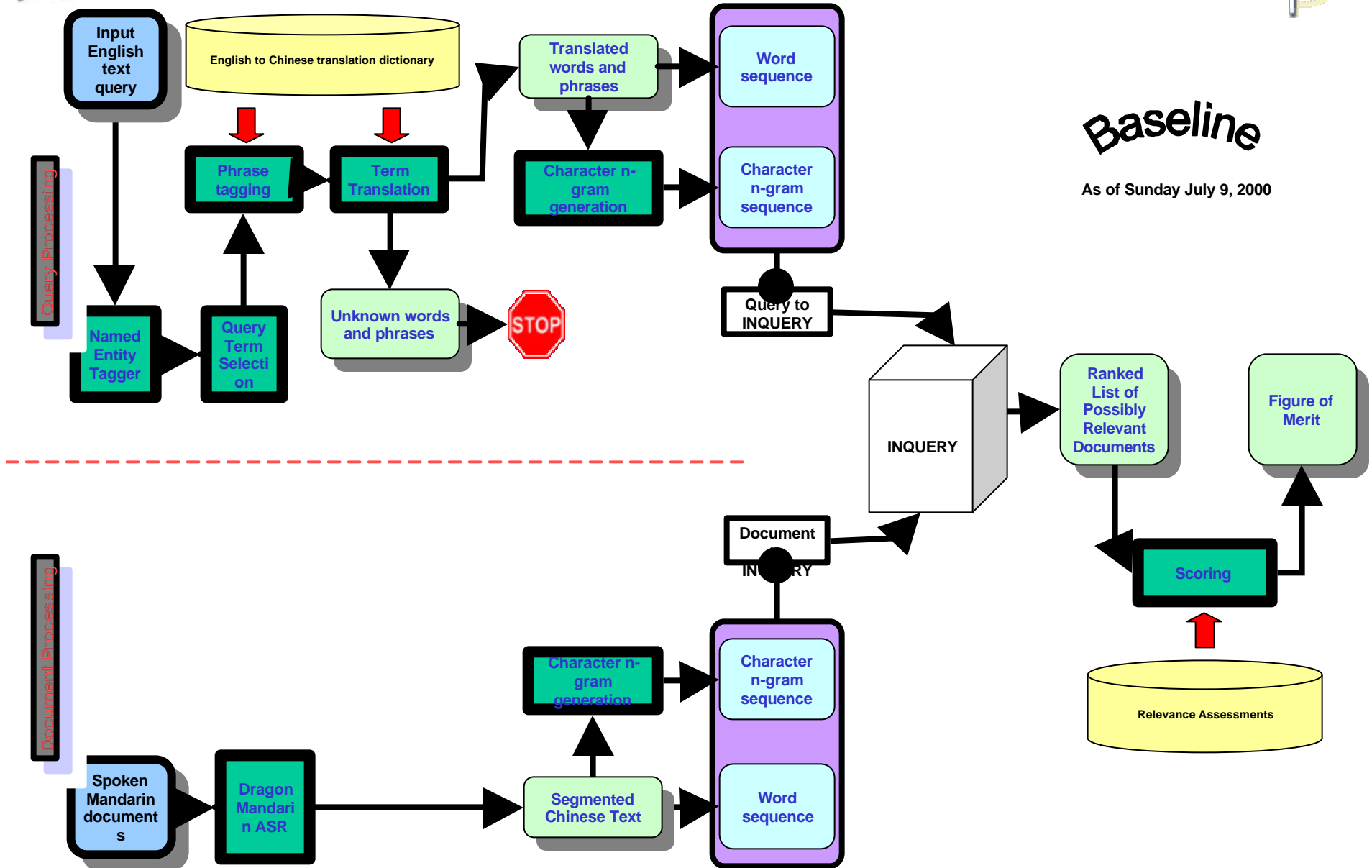
Mandarin-English Information: Investigation Translingual Speech Retrieval

Johns Hopkins University, Center for Language and Speech Processing, JHU/NSF Summer Workshop 2000

<<http://www.glue.umd.edu/~meiweb>>



MEI Team : Helen MENG (CUHK), Berlin CHEN (National Taiwan University), Erika GRAMS (Advanced Analytic Tools), Sanjeev KHUDANPUR (JHU/CLSP), Gina-Anne LEVOW (University of Maryland), Waik-Kit LO (CUHK), Douglas OARD (University of Maryland), Patrick SCHONE (Department of Defense), Karen TANG (Princeton University), Hsin-Min WANG (Academia Sinica), Jianqiang WANG (University of Maryland)



END