Topic Tracking at Maryland:
Lessons from the Johns Hopkins Mandarin-English Information (MEI) Project

Gina-Anne Levow and Douglas W.Oard
Institute for Advanced Computer Studies
University of Maryland, College Park

Roadmap

- MEI Overview (6 weeks in 5 minutes)
- MEI Results
- Adapting MEI to TDT
- TDT Results
- Conclusions

The 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

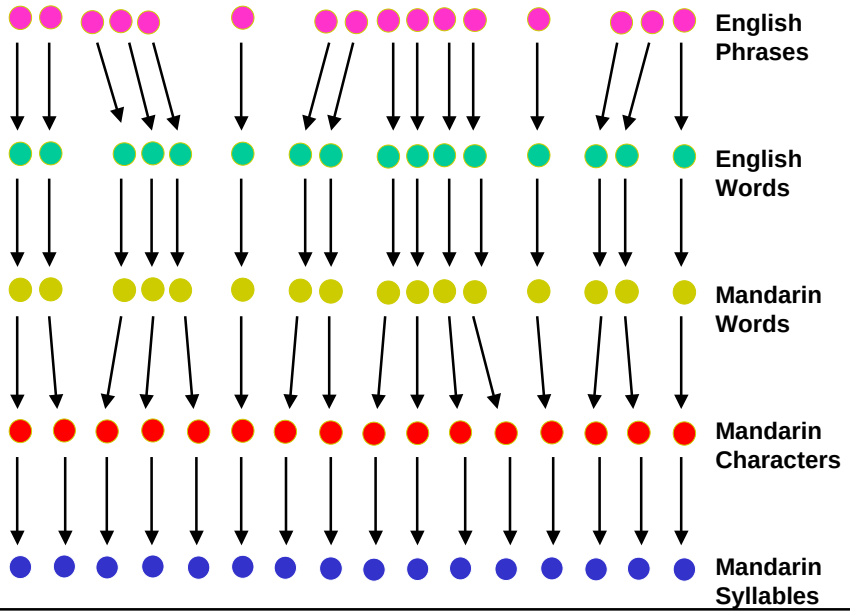
Berlin Chen	National Taiwan University
Wai-Kit Lo	Chinese University of Hong Kong
Karen Tang	Princeton University
Jianqiang Wang	University of Maryland

MEI: The Challenges

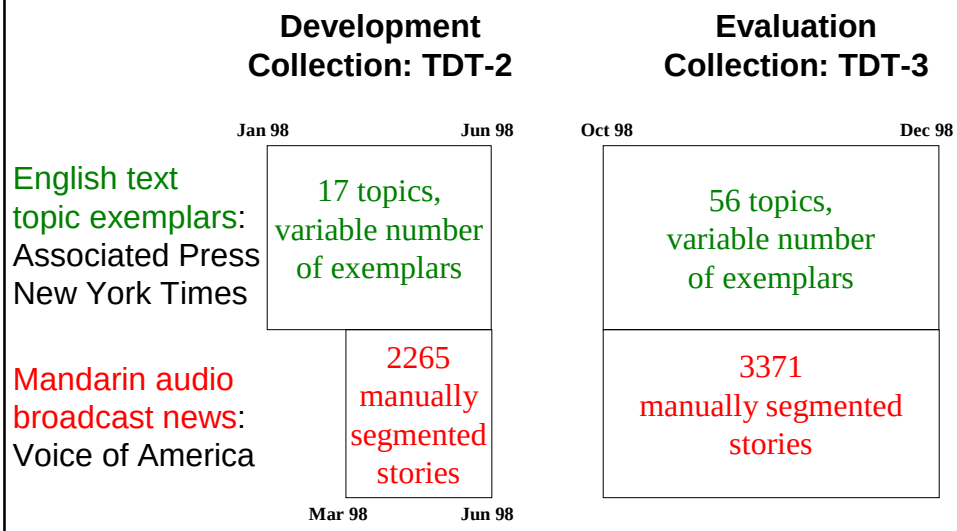
- Speech Recognition
  - Tokenization
  - Lexicon coverage
  - Selection among alternatives
- Translation
  - Tokenization
  - Lexicon coverage
  - Selection among alternatives

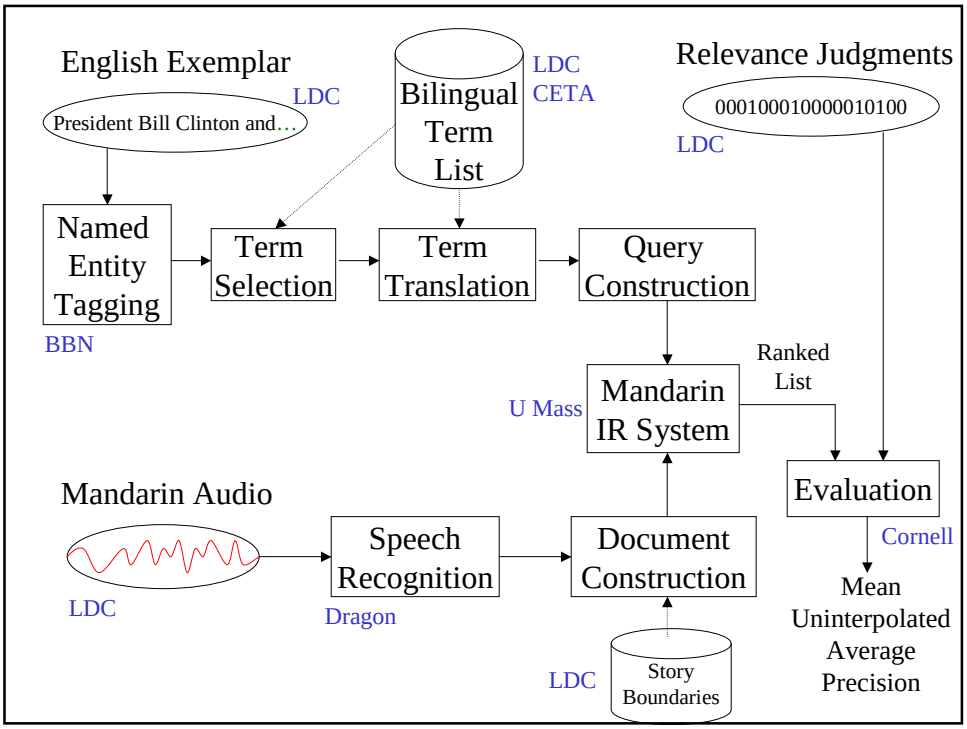
Different Problems

Term Granularity Options



MEI Evaluation Collections

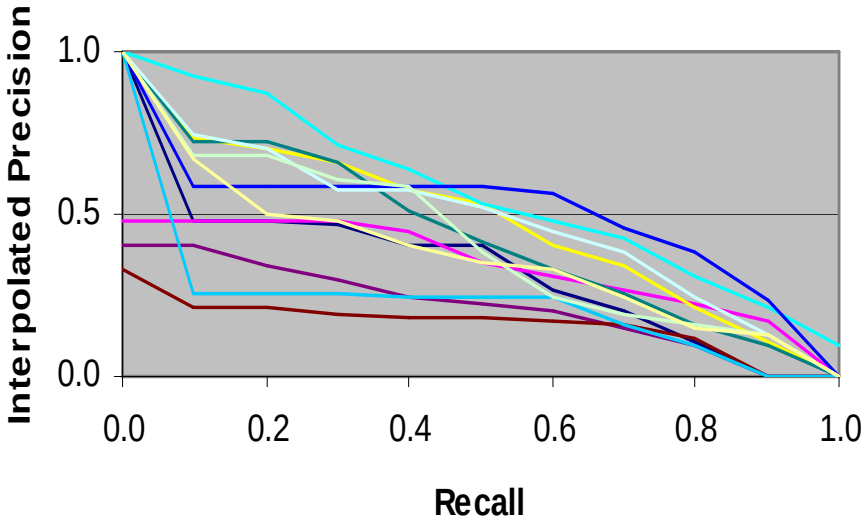




Query Translation

- Dictionary inversion for phrase translation
  - "Wall Street" "best interests" "human rights"
- Lemmatize remaining words if necessary
  - e.g. "televised" translates as "television
- χ^2 filtering for query term selection
  - Compared to an English background model

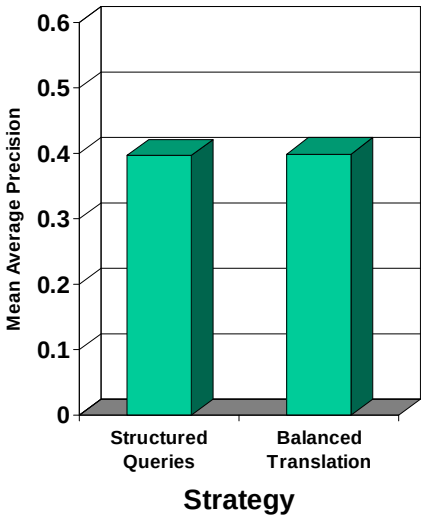
Evaluation Measure



Able to characterize variation across exemplars!

Balanced Translation Works Well

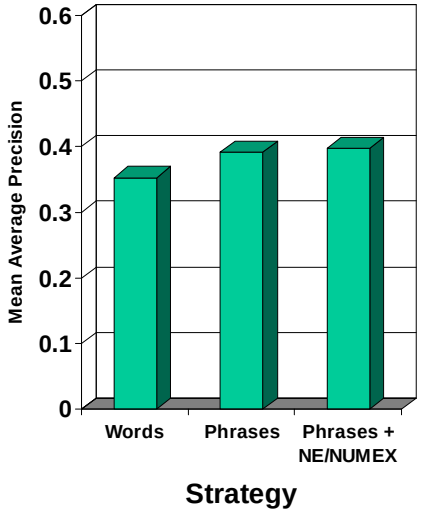
- Pirkola's structured queries
  - Treat translation alternatives as synonyms
  - Inquery #syn() operator
- Balanced translation
  - Distribute probability mass over translation alternatives
  - Inquery #sum() operator



TDT-2, phrase-based translation, word-based retrieval

Phrase Translation Beats Words

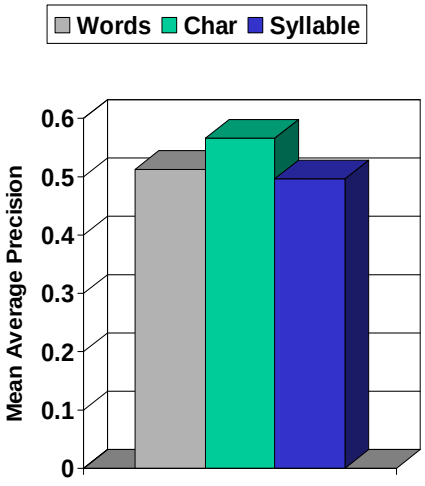
- Phrases beat words
- Three sources
  - Translation lexicon
  - Named entities
  - Numeric expressions



Condition: TDT-2, 12 exemplars, word-based retrieval

Character Bigram Indexing Wins

- Character bigrams are best
- Syllable bigrams do poorly



TDT-2, single NYT exemplar, manual translation

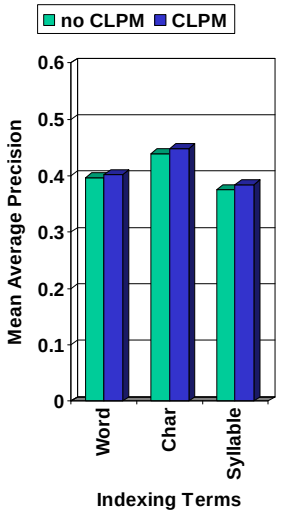
Untranslatable Terms

Terms	# (by token)	# (by type)
total	87,004	12,402
OOV	3,028	1,122

Term	Occurrences	
suharto		97
netanyahu	88	
starr	62	
arafat	50	
bjp	45	
vajpayee	44	
estrada		44
….		
hsu	19	
zemin	7	

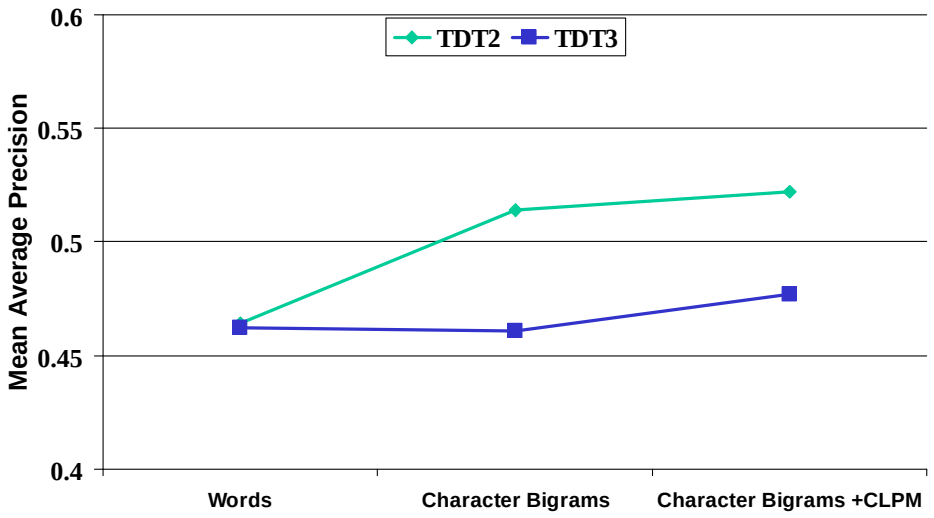
Cross-Language Phonetic Matching

- Small improvement
  - Not statistically significant
- Character bigrams are best
  - Form a unified index
    - Character and syllable bigrams
  - Translate words if possible
    - Then form character bigrams
  - Otherwise translate syllables
    - Then form syllable bigrams



TDT-2, phrase-based translation

MEI: Comparing Collections

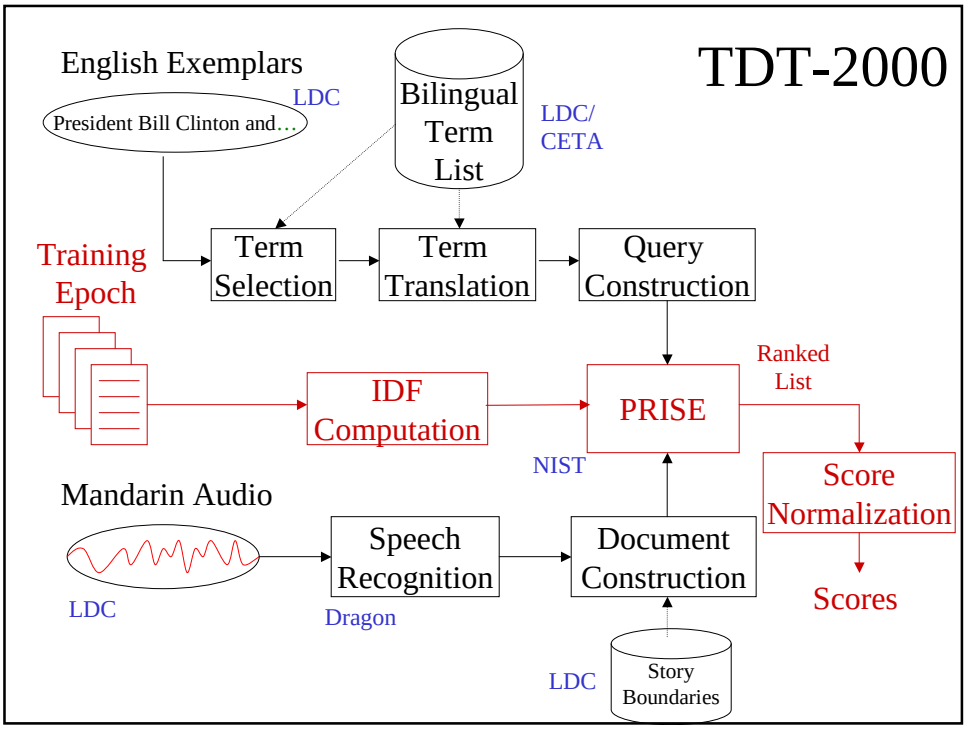


MEI Conclusions

- ASR Words
- Translation Phrases, Words, Lemmas, Syllables
- Indexing Character Bigrams

TDT-2000: What's New Since '99?

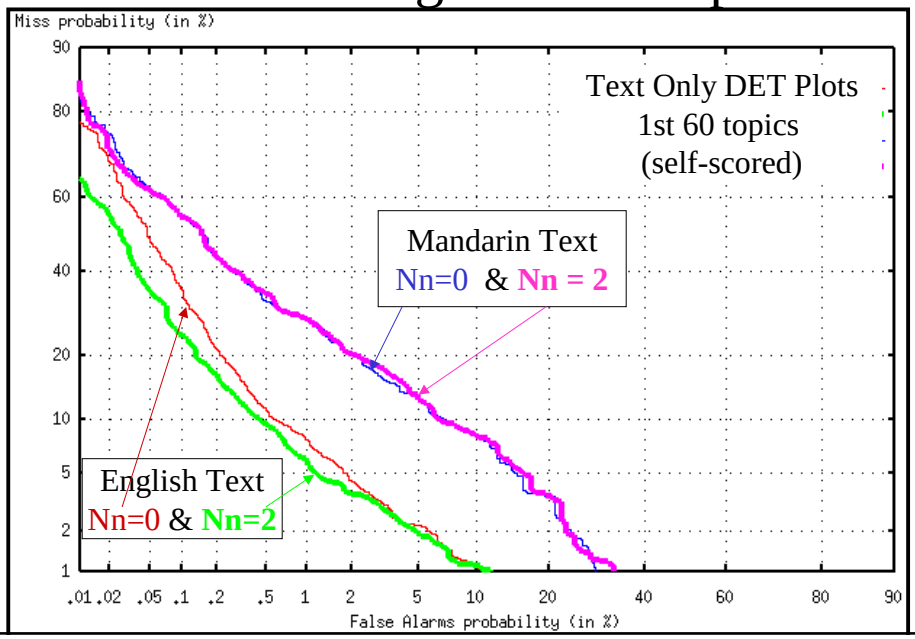
- Key ideas from MEI:
  - Dictionary inversion for phrase translation
  - Balanced translation
  - Post-translation resegmentation
- Adaptation to TDT:
  - Exploit negative exemplars
  - Improved Mandarin topic normalization
  - Round-robin balanced translation



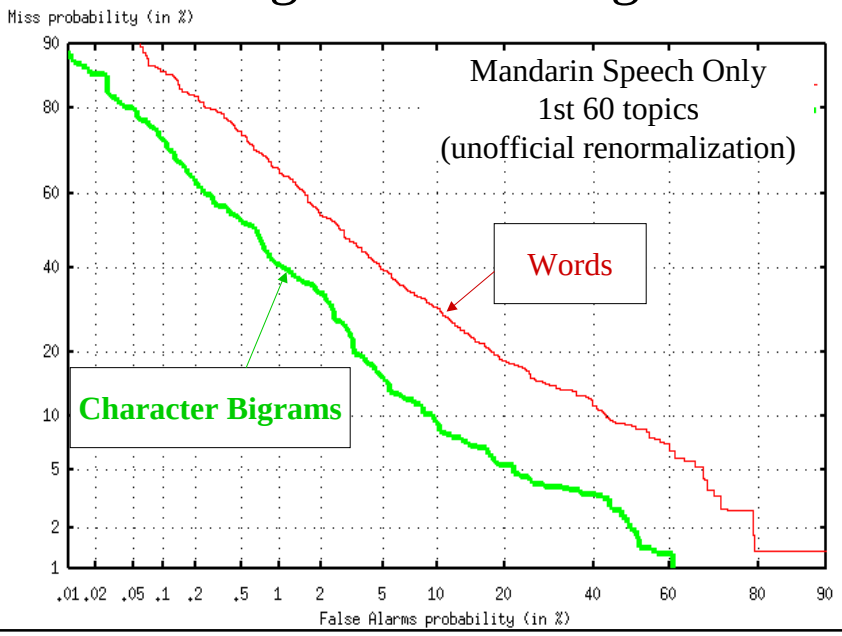
Topic Tracking Improvements

- Improved χ^2 filtering for query term selection
  - First compare to background model
  - Augment by comparison to negative exemplars
- Mandarin topic normalization (unofficial)
  - Language-specific strategy
    - Mandarin: Best single training epoch score
    - English: Average of exemplar scores
  - Recomputed Mandarin source normalization

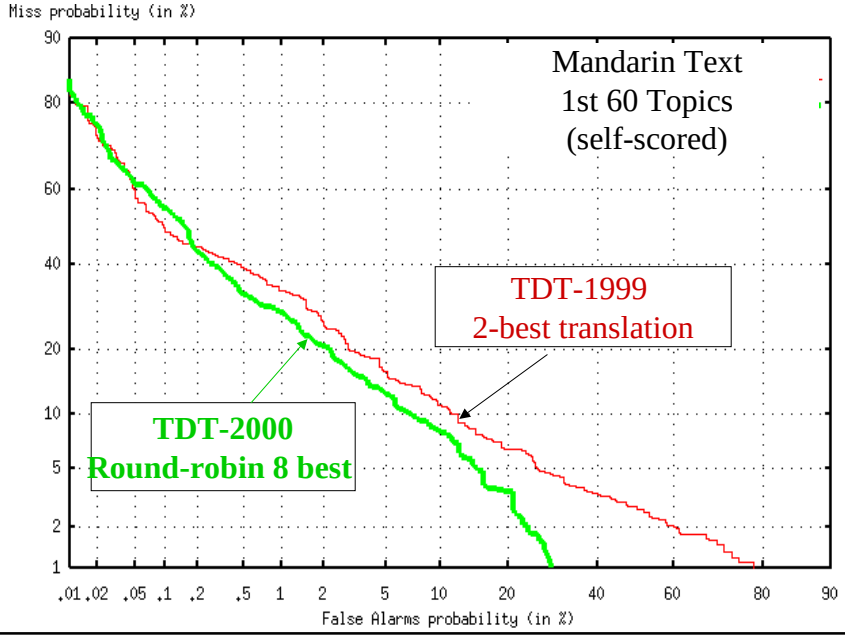
Effect of Negative Exemplars



Indexing Character Bigrams



Round Robin 8-Best Translation



Conclusions

- Top-8 round robin translation to Mandarin wins
  - Slightly outperforms top-2 translation to English
- Query translation is more efficient
  - Better suited to a stream of stories
- Match term extent to purpose
  - ASR, translation, indexing

Closing Thoughts

- Thanks to Jon and LDC !
- Normalization limits our insight
  - Need some way to see past it
- Availability of TDT-3 ground truth?