DARPA TIDES
MT Group Meeting

Marina del Rey

Jan 25, 2002

Alon Lavie, Stephan Vogel, Alex Waibel (CMU)

Ulrich Germann, Kevin Knight, Daniel Marcu (ISI)

Young-Suk Lee, Kishore Papineni, Salim Roukos (IBM)

Franz Josef Och (RWTH Aachen)

Moussa Bamba, Chris Cieri, Shudong Huang (LDC)
Florence Reeder (Mitre)

George Doddington (NIST)

Dec’01 Dry Run Chinese Review

- NIST-administered
- Testing on Xinhua, Zaobao, VOA data
- Training on HK News/Hansards, FBIS, dictionaries, Internet resources
- Humans beat machines on automatic scoring
- Data-driven research systems are competitive on Chinese “hard news” translation
- BLEU metric and NIST variations:
  - Statistical sensitivity experiments
  - Discussion of length penalty
  - Study of DARPA 94 data for automatic/manual score correlations
- Need to specify what “development test” data means!

Test Data Preparation

- LDC-produced
- 100 Chinese and Arabic hard news articles
- 10 human translations each
- Special instructions to translators
- Standards for formatting

- Human assessment of translations planned

June Chinese Evaluation: Training Data

- Available
  - LDC and public
    - C/E dict (new version out in March)
    - HK News, Hansards, Laws (not aligned!)
  - Publicly distributable (not yet at LDC)
    - HK News alignment (Ulrich Germann's ISI web site)
    - Chinese treebank translations (100K words)
      - Martha Palmer at UPenn
      - Alignments done by Ulrich Germann at ISI
  - Might be public if donated
    - HK Hansard alignment (BBN? Aachen?)
  - Will never be distributable
    - Old FBIS data (TIDES only)
    - Internet bilingual data
- Need to get all public data accessible through LDC!

Training Data (continued)

- Not yet available
  - UN data (50M words)
    - Legal problems
    - Alignment problems
  - New FBIS data (3m words?)
  - Xinhua bilingual data (possible 2m words): alignment
  - Various Internet bilingual data & dictionaries
  - Chem/bio dictionaries, State Dept, CDC

June Chinese Evaluation: Testing

- Testing
  - Chinese news plus DARPA discretion
- Administration & Scoring
  - NIST web page, mailing lists, proselytizing
  - BLEU/NIST scoring
  - Human assessment

June Chinese Evaluation: Tracks

- Open track
  - Any resources/methods ok
  - No web crawling after March 15
- Large common data track
  - No bilingual texts except LDC-available ones
  - No bilingual dicts except LDC-available ones
  - Any monolingual texts you want, tokenizers, etc, etc
- Small common data track
  - No bilingual texts except 100K-word Chinese treebank texts
  - Can’t use trees from treebank
  - 10K dictionary to be released by S. Vogel in March
  - No corpus-trained tools like trained LDC tokenizer
  - Any English text, treebanks, etc, okay to use

June Chinese Evaluation: Timeline

- March 15
  - no more web-crawling by participants
  - test data collection begins!
- June 3: test data out
- June 14: submissions due
- June 21: raw results distributed
- July 22-23: workshop

June Arabic Evaluation

- No dry run for Arabic – the Chinese dry run should suffice for debugging the eval procedure
- Training
  - LDC aiming at 100K word POS tagged (when?)
  - LDC aiming at morph analyzer (when?)
  - LDC aiming at Ummah 100K word bilingual text (when?)
  - Koran used at IBM
  - UN data available at some sites (50M words)
- Testing – same as Chinese
- Administration & Scoring – same as Chinese
- Tracks – open track only, for first go-round
- Timeline – same as Chinese