HLT Evaluation: Role of Data Centers

Christopher Cieri
ccieri@ldc.upenn.edu

University of Pennsylvania
Linguistic Data Consortium and Department of Linguistics
3600 Market Street, Philadelphia, PA 19104 U.S.A.

www.ldc.upenn.edu

Common Task Evaluation

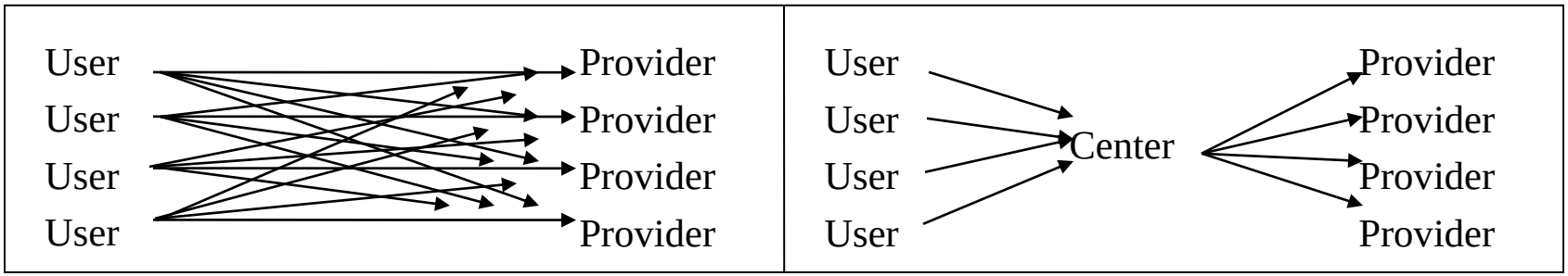
- Research management paradigm used commonly in DARPA programs
  - Sponsors, R&D Sites, Data, Evaluation Sites
    - Sponsors: DARPA, ARDA, NRL
    - R&D Sites: **international** non-profit, commercial, government sites
      - TREC, Cambridge and LIMSI in EARS
    - Data and Data Coordination often but not necessarily by LDC
      - MITE for SPINE
    - Evaluation often but not necessarily by NIST
      - NRL for SPINE, BAE for GALE
  - Tasks, schedule, evaluation metric, data needs defined jointly
  - Tasks include required; may include optional
  - Focus tends to be on technology. Evaluation tends to be objective, metrics based
  - Evaluation metrics, data, systems descriptions and research papers shared
  - Competition/cooperation vary with program goals/incentives
    - Martin's comment!

Plan

- Attend planning and technical meetings
  - Understand research needs; support in data plan
  - Data Plan coordinated with Evaluation Plan
- Communicate issues regarding data creation, annotation, distribution
  - Manage expectations
  - Explain limitations on human annotation
  - Maintain timeline
- Budget
  - Compromise, make tough decisions
  - Allocate funds for evaluation data first, remainder for training, development/text
- Maintain progress orientation

Collect, Annotate

- Collect directly or subcontract
- Identify Data Epochs
  - Training, Development/Test, Evaluation
- Coordinate collection to data needs
- Quality Control on ongoing basis
  - Inter-annotator agreement versus quality
  - precision, recall, dual annotation, adjudication
- Acquire intellectual property rights upfront
  - manages risk on expensive annotated data



Distribute

- User Agreements
  - Members Agreements, Non-Member Licenses
  - Evaluation Agreement
    - rights dependent upon participation
- Distribution within and outside program
  - Reserving attracts participants to program.
  - Sharing attracts participants to research.
  - All resources eventually shared generally
  - Evaluation data reserved until replaced
  - Progress Sets? Mothballed Systems?

Evaluation Data Sets

LDC2004T15 2000 Communicator Dialogue Act Tagged
LDC2004T16 2001 Communicator Dialogue Act Tagged
LDC2004S04 2002 NIST Speaker Recognition Evaluation
LDC2004S11 2002 Rich Transcription Broadcast News and Conversational Telephone Speech
LDC2004S09 NIST Meeting Pilot Corpus Speech
LDC2004T13 NIST Meeting Pilot Corpus Transcripts and Metadata
LDC2004S07 Switchboard Cellular Part 2 Audio
LDC2003T03 1997 HUB5 German Transcripts
LDC2003T04 1997 HUB5 Spanish Transcripts
LDC2003T02 1998 HUB5 English Transcripts
LDC2003S01 2001 Communicator Evaluation
LDC2002S11 1997 HUB4 English Evaluation Speech and Transcripts
LDC2002S22 1997 HUB5 Arabic Evaluation
LDC2002T39 1997 HUB5 Arabic Transcripts
LDC2002S24 1997 HUB5 German Evaluation
LDC2002S25 1997 HUB5 Spanish Evaluation
LDC2002S10 1998 HUB5 English Evaluation
LDC2002S56 2000 Communicator Evaluation
LDC2002S13 2001 HUB5 English Evaluation
LDC2002S12 2001 HUB5 Mandarin Evaluation
LDC2002S34 2001 NIST Speaker Recognition Evaluation Corpus
LDC2002S06 Switchboard-2 Phase III Audio
LDC2001S91 1997 HUB-4 Broadcast News Evaluation Non English Test Material
LDC2001S97 2000 NIST Speaker Recognition Evaluation
LDC2001S04 Speech in Noisy Environments (SPINE2) Part 1 Audio
LDC2001T05 Speech in Noisy Environments (SPINE2) Part 1 Transcripts
LDC2001S06 Speech in Noisy Environments (SPINE2) Part 2 Audio
LDC2001T07 Speech in Noisy Environments (SPINE2) Part 2 Transcripts
LDC2001S08 Speech in Noisy Environments (SPINE2) Part 3 Audio
LDC2001T09 Speech in Noisy Environments (SPINE2) Part 3 Transcripts
LDC2001S99 Speech in Noisy Environments 1 (SPINE1 CODED) Coded Audio
LDC2001S13 Switchboard Cellular Part 1 Audio
LDC2001S15 Switchboard Cellular Part 1 Transcribed Audio
LDC2001T14 Switchboard Cellular Part 1 Transcription
LDC99S79 Switchboard-2 Phase II
LDC99S84 TDT2 English Audio
LDC98S71 1997 English Broadcast News Speech (Hub-4)
LDC98T28 1997 English Broadcast News Transcripts (Hub-4)
LDC98S73 1997 Mandarin Broadcast News Speech (Hub-4NE)
LDC98T24 1997 Mandarin Broadcast News Transcripts (Hub-4NE)
LDC98S74 1997 Spanish Broadcast News Speech (Hub-4NE)
LDC98T29 1997 Spanish Broadcast News Transcripts (Hub-4NE)
LDC98S76 1998 Speaker Recognition Benchmark
LDC98S69 Hub-5 Mandarin Telephone Speech Corpus
LDC98T26 Hub-5 Mandarin Transcripts
LDC98S70 Hub-5 Spanish Telephone Speech Corpus
LDC98T27 Hub-5 Spanish Transcripts
LDC98S75 Switchboard-2 Phase 1
LDC97S66 1996 English Broadcast News Dev and Eval (Hub-4)
LDC97S44 1996 English Broadcast News Speech (Hub-4)
LDC97T22 1996 English Broadcast News Transcripts (Hub-4)
LDC97S62 SWITCHBOARD-1 Release 2
LDC96S61 1996 Speaker Recognition Benchmark
LDC96S36 Boston University Radio Speech Corpus
LDC96S33 CSR-IV Hub 3
LDC96S31 CSR-IV Hub 4
LDC95S26 ATIS3 Test Data
LDC94S16 YOHO Speaker Verification
LDC93S3B Resource Management RM1 2.0
LDC93S3C Resource Management RM2 2.0
LDC93T3A TIPSTER Complete
LDC93T3B TIPSTER Volume 1
LDC93T3C TIPSTER Volume 2
LDC93T3D TIPSTER Volume 3
LDC2001S93 TDT2 Mandarin Audio Corpus
LDC2001T57 TDT2 Multilanguage Text Version 4.0
LDC2001S94 TDT3 English Audio
LDC2001S95 TDT3 Mandarin Audio
LDC2001T58 TDT3 Multilanguage Text Version 2.0
LDC2000S86 1998 HUB-4 Broadcast News Evaluation English Test Material
LDC2000S88 1999 HUB-4 Broadcast News Evaluation English Test Material
LDC2000S96 Speech in Noisy Environments (SPINE) Evaluation Audio
LDC2000T54 Speech in Noisy Environments (SPINE) Evaluation Transcripts
LDC2000S87 Speech in Noisy Environments (SPINE) Training Audio
LDC2000T49 Speech in Noisy Environments (SPINE) Training Transcripts
LDC2000S92 TDT2 Careful Transcription Audio
LDC2000T44 TDT2 Careful Transcription Text
LDC2000T52 TREC Mandarin
LDC2000T51 TREC Spanish
LDC99S80 1997 Speaker Recognition Benchmark
LDC99S81 1999 Speaker Recognition Benchmark