Resources -- New and Forthcoming -- from LDC

Dave Graff
Stephanie Strassel
Christopher Cieri

Overview

- **Broadcast News Collection & Transcription**
  - TIDES VOA Data, VOA Czech & Korean, TDT2 English, 1999 Hub-4 Test Sets
- **Telephone Speech Collection, Transcription and Lexicography**
  - Korean, Russian, Farsi, Switchboard Cellular
  - More Switchboard Cellular
- **NRL SPeech In Noisy Environments**
- **Other Projects in Annotation, Architecture, Tools**
  - TDT - annotate 60 new topics from TDT-3 corpus
  - ACE - entity recognition annotation along with other sites
  - ATLAS - architecture for linguistic analysis
  - Talkbank - standards, tools to annotate communicative interactions

Broadcast Transcription

Completed Projects to be published in 2000

- **TDT2 English**
  - 10 hours, transcribed to Hub4 specifications minus background conditions in 1999
- **VOA Czech**
  - ~ 30 hours, transcribed for 1999 JHU Summer Workshop by Charles University, Prague
- **1999 Hub4 English and Mandarin Test Sets**
  - including background conditions

New Projects for 2000

- **VOA Korean**
  - 10 hour sample, under discussion with MIT-LL
- **TDT2 English (for NIST)**
  - another 10 hours sample for NIST SDR

Broadcast Collection

- **Voice of America broadcasts**
  - 40-50 languages
  - from 15 minutes to 1 hour per day
  - low density languages represented, languages change over time
  - partially parallel - stories translated from in-house new bureau or written specifically for the branch
- **LDC began sampling in October 1999**
  - up to 8 channels recorded simultaneously, selected from 48 available (many channels provide multiple languages each day)
  - 15 - 22 broadcast hours digitized daily
  - program boundaries labeled manually
  - data is sorted by language, archived and moved to near-line tape storage
- **All VOA broadcast languages are now present in the archive (though some have started only recently)**

CallFriend

- **CallFriend transcription completed, 3 new languages**
  - Korean
    - » using CallFriend calls collected in 1996
    - » 100 calls * 15 minutes transcribed in native orthography (KSC-5601)
  - Russian
    - » 140, 30-minute calls newly collected in 1999
    - » 80 calls * 15 minutes transcribed in native orthography
  - Farsi (Thanks to BBN)
    - » using CallFriend calls collected in 1996
    - » 60 calls * 10 minutes transcribed in romanized orthography
    - » will be expanded in 2000 to 100 calls * 15 minutes transcribed

- **New lexicons to cover transcripts in Russian, Korean, Farsi nearly finished.**

Switchboard Cellular

- **Pilot collection just ending**
  - goal 190 speakers, >1000 calls
  - focus on GSM cellular network
  - all calls audited
- **Additional collection to begin later this year**
  - goal 210 speakers, 10 sides each
  - wider variety in cellular network
- **Transcription to begin in 2000**
  - 250 sides * 5 minutes
  - lexical enhancements
  - maximize range of speakers represented

NRL SPINE

- Two-channel, semi-scripted, cooperative task interactions in a "battleship" game scenario
- Multiple sessions by pairs of speakers using varied signal conditions
  - handsets, channels, vocoders, background noises
- Limited vocabulary, including "diagnostic-rhyme test" word pairs (e.g. "fame"-"same")
- Initial transcription by professional service (FDCH)
- Final time-alignment and QC of transcripts by LDC

- First training set comprises ~10 hours of speech
- Test data to follow later this summer

1999 Hub4 Test Sets

- Professional services did initial Eng/Mand. text for 10-hour test pools
- LDC annotators applied time alignments for story boundaries
- NIST selected test sets from initial transcripts and mark-up
- LDC then
  - added turn- and phrase-level time stamps, overlap and disfluency
  - used HTK forced alignment to get word-level time marks
  - did third manual labeling pass to identify background conditions (observing word boundaries)
  - folded background labels and time-stamps into transcripts by matching to the nearest ASR-aligned word boundary