Linguistic Resources for Meeting Recognition

Meghan Glenn, Stephanie Strassel
Linguistic Data Consortium

{mlglenn, strassel@ldc.upenn.edu}

http://projects.ldc.upenn.edu/Transcription/NISTMeet
Overview

• Training data distribution
• New corpus creation
  – Transcription team
  – NIST Phase 2 Corpus Part 2
    • Quick transcription
  – Conference room test data
    • Careful transcription
    • Quality control
• Infrastructure
  – XTrans Toolkit
    • Existing features for meetings
    • Future features for meetings
• Data collection opportunities
## RT-09 Training Data
provided by LDC

<table>
<thead>
<tr>
<th>Title</th>
<th>Speech</th>
<th>Transcripts</th>
<th>Volume</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher English Part 1</td>
<td>LDC2004S13</td>
<td>LDC2004T19</td>
<td>750+ hours</td>
<td>CTS</td>
</tr>
<tr>
<td>Fisher English Part 2</td>
<td>LDC2005S13</td>
<td>LDC2005T19</td>
<td>750+ hours</td>
<td>CTS</td>
</tr>
<tr>
<td>ICSI Meeting Corpus</td>
<td>LDC2004S02</td>
<td>LDC2004T04</td>
<td>72 hours</td>
<td>Meeting</td>
</tr>
<tr>
<td>ISL Meeting Corpus</td>
<td>LDC2004S05</td>
<td>LDC2004T10</td>
<td>10 hours</td>
<td>Meeting</td>
</tr>
<tr>
<td>NIST Meeting Pilot Corpus</td>
<td>LDC2004S09</td>
<td>LDC2004T13</td>
<td>13 hours</td>
<td>Meeting</td>
</tr>
<tr>
<td>RT-04S Dev-Eval Meeting Room Data</td>
<td>LDC2005S09</td>
<td>LDC2005S09</td>
<td>14.5 hours</td>
<td>Meeting</td>
</tr>
<tr>
<td>RT-06 Spring Meeting Speech Evaluation Data</td>
<td>LDC2006E16</td>
<td></td>
<td>3 hours</td>
<td>Meeting</td>
</tr>
<tr>
<td>TDT4 Multilingual Broadcast News Corpus</td>
<td>LDC2005S11</td>
<td>LDC2005T16</td>
<td>300+ hours</td>
<td>BN</td>
</tr>
</tbody>
</table>
RT-09 Transcriber team

- Native English speakers
  - Diverse backgrounds
    - Teachers, musicians, linguists
- Previous transcription experience
  - Quick transcription method
  - Telephone speech
  - Meeting recordings
    - Most new to meeting transcription task
    - Previous meeting transcription efforts used the same transcribers from eval to eval
  - Very little previous experience with careful transcription approach
NIST Phase 2 Corpus

• Data profile
  – 5 hours
  – 3 - 5 speakers per session
  – Sessions range from 33-106 minutes
  – Primarily native English speakers
  – Topic content
    • Business meetings
    • Paper and presentation reviews

• Quick transcription approach
  – IHM recordings
  – Manual segmentation + transcription in single pass
    • Targets content words
    • Transcribers listen to segments once or twice
    • Markup of acronyms and spoken letters
    • No markup of filled pauses or proper nouns
    • Optional (additional) modification of segmentation
  – Quality control to resolve and standardize proper nouns, “uncertain transcription”
RT-09 Test data overview

- Conference room data
  - Seven meeting session excerpts
  - 4 - 11 speakers per session
    - Native and non-native
  - 19-30 minutes each
  - Three contributing sites
  - Multiple recording conditions for each session
  - Varied Topic content
    - Product/new technology design
    - Data collection
    - Economic discussion
    - High school arts program discussion
    - Baby shower planning
    - Small business owners workshop
RT-09 Test transcription (1)

• Careful transcription
  – IHM recordings, one speaker per channel
  – *Stage 1: Manual Segmentation*
    • Segments are breath groups
      – Average 3-8 seconds, primarily for ease of transcription
      – ~10 ms padding at edges of segment boundaries
      – Turn-taking structure
    • 1 X RT
Stage 2: Verbatim Transcription

- Slow, very careful orthographic transcription
- No time limit
- Speaker and background noise
  - Vocalized noise – limited to 5 sounds
    » Ignore consistent heavy breathing
Stage 3: Transcription Verification and Markup

- Add markup for filled pauses, proper names etc.
- Verify segmentation & transcription accuracy
- Revisit difficult sections
  - Acronyms, technical terms, proper nouns

3 X RT
Quality Control (CTR)

- After transcription of all speaker channels in a meeting
- Use merged IHM recordings or distant microphone recording
- Focus on
  - Transcription & segmentation accuracy, completeness
  - Speaker ID consistency
  - Consistency, accuracy of names, acronyms, terminology
  - Examine silence (untranscribed) regions for missed speech using customized tool functions
  - Markup consistency
  - Final spell check
  - Expand contractions
  - Export to CTS (.txt) format
Unique Challenges

• Multiple speakers
  – Overlapping speech
  – Asides

• Meeting content
  – Acronyms
    • Example: “WIIFM”
  – Project discussion groups
  – Role playing meetings

• Meeting spaces
  – Ambient noise

• Varying levels of speaker participation

• No video access
  – In the works for future versions of XTrans
  – Improve speaker ID, especially for ambient speakers
Infrastructure

XTrans toolkit

– Generalized speech annotation tool
– Multi-platform, multi-lingual, multi-domain
– Based on QT, implemented in Python and C++
– Component-based, reconfigurable for new tasks
– Extensible to other tools
  • QCTool for translation quality control
– Built-in support for common LDC tasks
  • Quick and careful transcription
  • Structural spoken metadata annotation
  • Meetings, conversational telephone and broadcast speech

• Linux version available on LDC website
  – http://projects ldc.upenn.edu/gale/Transcription/download_xtrans-linux-latest.php
XTrans Features

• User-friendly GUI
  – All commands can be issued from keyboard or from mouse
  – Keyboard-only is much faster
  – User-configurable keybindings for common tasks

• Bi-directional text input
  – Critical for languages like Arabic

• SpeakerID verification functions include
  – LRS: Listen to a random segment from this speaker to verify voice
  – LAS: Listen to all segments from this speaker in the file

• Waveform display/playback components
  – QWave, based on QT
    • Variable speed playback
    • Relative volume control for individual channel
    • Amplitude control

• Inter-gap playback
  – LAG: Listen to the unsegmented audio "gaps" (helpful for doing quality control, to catch unsegmented speech)
XTrans Features (2)

- Virtual speaker channels
  - One VSC per *speaker*, not per audio channel
  - Enables easy handling of overlapping speech in single-channel audio, ideal for meeting recording quality control
XTrans Features (3)

- Fluid single vs. multiple speaker focus
  - Arbitrary number of audio channels can be loaded at once
  - Toggle between multiple playback functions
    - Merged IHM
    - Multiple individual IHMs
    - Single IHM for one speaker
    - Any channel can be muted
  - Toggle between merged, multi-speaker transcript view and single-speaker view
    - Use complete transcript for context
- Waveform markup display makes speaker interaction obvious
- Easy creation/modification of configuration files makes transcription more efficient
Single Speaker Focus

- Transcript and audio linked – click in transcript, focus on that region in the audio
- Focus on one speaker
- Mute other speaker channels
Multi-Speaker Focus

- Speakers interleaved in transcript
- Listen to all speakers at once
Impact of XTrans

• Quality control
  – Better integrated into tool itself rather than stand-alone post-process
  – Customized features support QC at all stages
  – Adjudication module for comparing two versions transcripts

• Real time transcription rates
  – RT-05: over 65 x real-time for QTR
  – RT-07: approximately 50 x real-time for CTR
  – RT-09: approximately 40 x real-time for CTR
Future directions

• XTrans development
  – MP3 audio capability
  – Video playback capability
    • Easier speaker ID
    • Easier meeting “contextualization”
  – Integrate additional annotation functions
    • Currently stand-alone modules
      – Contraction expansion
    • New text display component for rich disfluency annotation
      – Will enable transcript correction during annotation tasks
      – Dialect annotation
  – Better non-English (non-Roman) input methods
    • Currently rely on SCIM and other external protocols

• Data collection
  – LDC has some meeting collection capability
    • Conference room domain
      – Interview sessions
      – Small group sessions
    • Possible lecture room opportunities
      – Portable collection platform