Corpus Creation for Disfluency Research

Stephanie Strassel
Linguistic Data Consortium
{strassel@ldc.upenn.edu}
The Linguistic Data Consortium supports linguistic research, education and technology development by creating and sharing linguistic resources: data, tools and standards

Data
- More than 16,000 copies of more than 230 corpora distributed to more than 1300 organizations
  - Publish 25+ corpora/year to members; most available to non-members
  - Plus dozens of “e-corpora” to provide training and evaluation data for sponsored common task evaluations
- Sponsorship from funded projects, community or LDC initiatives
- Conversation, interview, task-oriented dialog, broadcast radio & television, read speech, news text, parallel text & lexicons in many languages
- Video, speech and text annotation in many languages including
  - Transcription, POS tagging, morphology tagging, treebanking
  - Entity, relation & event tagging, topic relevance tagging for information retrieval
  - Sociolinguistic variation, lexicons, gesture
  - “Metadata tagging” – including disfluencies
- Customized annotation and corpus development tools using Annotation Graph model
Introduction

• Staff
  – 37 fulltime staff covering external relations, data collection and creation, research and development
  – 60+ part-time staff for annotation, technical and admin support
    • Annotator backgrounds vary
    • Linguistics training sometimes not necessary or even desirable

• Evolutionary Paths
  – Demands: more data, wider variety of languages, new data modes and types, increasingly complex annotation, broader range of communities to serve
  – Solutions: research best practices, provide tools, offer value added services, reuse resources, link research communities
DARPA EARS Program
(Effective, Affordable, Reusable Speech-to-Text)

Enables development of core speech-to-text technology to produce rich, highly accurate automatic speech recognition output in a range of languages and speaking styles.

Aggressive program goals target *substantial* improvements on current technology in English, Chinese and Arabic; in conversational telephone speech and broadcast news.

DiSS ’03 Workshop
MDE Task

• “Metadata” Extraction
  – Detect & characterize certain linguistic features, in order to
    • Output cleaned-up, structured transcript
    • With ultimate goal of improved transcript readability

• Primary Metadata Features
  – Fillers
    • Filled pause, discourse marker, optional editing terms
      – Asides & parentheticals
  – Edit Disfluencies (or speech repairs)
    • Repetitions, revisions, restarts, complex
  – SUs (“semantic” units)
    • Statement, question, backchannel, incomplete
      –Clausal and coordinating internal SUs

• Task defined with “clean-up” in mind
well um i work in a fac- or a building that’s that’s not really it well it’s on the campus of the main company but it’s a little bit you know separated and um it’s mo- it’s mainly a factory environment

Example from Switchboard
...and not an atypical one
well um I work in a fac- or a building that's that's not really it well it's on the campus of the main company but it's a little bit you know separated and um it's mo- it's mainly a factory environment
well um i work in a factory or a building that's not really it well it's on the campus of the main company but it's a little bit you know separated and um it's mainly a factory environment
well um I work in a fac-or a building | that's not really it well it's on the campus of the main company | but it's a little bit you know separated | and um it's mo- it's mainly a factory environment |
well um I work in a fac- or a building. that's
that's not really it well It's on the campus of
the main company, but it's a little bit you
know separated. And um it's mo- it's mainly
a factory environment.
<Joe_Smith>
I work in a building.
It's on the campus of the main company, but it's a little bit separated.
And it's mainly a factory environment.
......
Full Metadata Task: Edit Disfluencies

- Identify
  - Original utterance (reparandum)
  - Interruption point
  - Optional editing term (interregnum)
  - Correction (repair)

- Classify
  - Repetition
    
    [He-] * he's really out of line, or at least that's what I was told

  - Revision
    
    Fifty-six residents were [killed] * er injured rather.

  - Restart-Keep: content should be preserved in cleaned-up transcript
    
    [I happen to live not too far away]K * well, I’ve actually worked for the company that has been blamed for the Challenger disaster.

  - Restart-Discard: content should be removed in cleaned-up transcript
    
    [It's also]D * I used to live in Georgia.

  - Complex (multiple, nested edits)
    
    I'm sure [the] * that [the uh] * the staff learn what's normal...
Task a moving target
  – Especially problematic with annotation team approach and aggressive schedule, data demands

Low consistency, very slow

Errors in underlying transcripts

Spending a lot of time on rare constructions

DiSS ’03 Workshop
DiSS '03 Workshop

Defining the Metadata
Task: Solution

- Tag the **depod**: Deletable portion of disfluency
  - Equivalent to the original/reparandum portion
- Do not specifically label
  - Edit type
  - Corrected portion
- Label all interruption points
  - Automated at right edge of depod
- Collapse all nested, serial edits into single depod with multiple interruption points
- “Difficult decision”, “no annotation”, “bad transcription” labels

[It’s * this is like only like the third or fourth time I’ve * I ne- * I’m real bad about] * I never make the phone calls
SimpleMDE Task: Implications

• Provides baseline annotation
  – Does not model everything
  – Further detail possible at later stages
• Enables high volume data production
  – On aggressive schedule
• Removes uncertainty from task
  – Even for non-expert annotators
• Encourages better inter-annotator agreement
  – Important given annotation team approach
## MDE Data Overview

<table>
<thead>
<tr>
<th>Task</th>
<th>Full Metadata Task</th>
<th>Simple Metadata Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase</td>
<td>Startup</td>
<td>Moving Target</td>
</tr>
<tr>
<td>Corpus</td>
<td>Micro-corpus</td>
<td>Mini-Train, DevTest</td>
</tr>
<tr>
<td>Data in minutes</td>
<td>6 minutes</td>
<td>12.5 hours</td>
</tr>
</tbody>
</table>

- Broadcast news: recent data from Hub-4 Corpus
  - Single channel, multiple speakers (overlapping speech)
  - Fewer edit disfluencies; many difficult SUs
- Conversational Telephone Speech: from Switchboard and Fisher
  - Two channels, two speakers
  - Subset of data drawn from Penn Treebank-3
    - Includes Meteer-style disfluency annotation, POS, Treebank
  - *Many* edit disfluencies, fillers
  - SUs somewhat easier to detect and characterize
SimpleMDE Annotation Tool

• Annotation Graph model
  – Infrastructure for annotation tools and data format
• Standoff markup, XML
  – Each feature a separate annotation layer
• Multi-platform, multi-lingual
• Written in Python
• Freely available www.ldc.upenn.edu/Projects/MDE
• User features
  – Audio, transcript in sync
  – Fillers are pre-tagged
  – Displays annotation with color, underline
  – Monitors annotation for common errors
  – User can view each annotation layer (type) separately or integrated for QC
  – User can view cleaned-up transcripts for QC
SimpleMDE Annotation Tool

Usage:
- Swipe over text
- Play audio (one or both channels)
- Add annotation
- Key- and mouse-bindings for common tasks

DiSS ’03 Workshop
Quality Control

• Annotator selection and training
  – Do careful transcription as well, to understand context

• Searchable annotator-created web guidelines
  – Many additional examples
  – Includes log of questions and resolutions

• Customized annotation tool
  – With custom views for second passing, QC, adjudication
  – Validation and automatic scans for common errors

• Second pass over every file
  – Performed by independent annotator
  – Each annotation type reviewed separately
    • Can hide or display other annotation layers as needed
  – All difficult decisions reviewed again by team leader

• 10% of data dually annotated
  – By independent annotator
  – Adjudication and resolution of discrepancies

• All QC results feed back into annotator training & guidelines
SimpleMDE Adjudication Tool

Details of annotation discrepancies
Conclusions & Future Work

• Current corpus
  – Currently available to EARS community only
    • After evaluation, regular publication
  – Non-expert annotation team approach working well
    • CTS: <20x real time for two complete passes
    • BN: <15x real time for two complete passes
  – Inter-annotator agreement good
    • Now ~97% agreement for depod, IP, filler detection/characterization

• Likely future directions
  – Additional SimpleMDE training data
  – Richer (Full MDE?) annotation for subset of data
  – Expand to Mandarin Chinese and Arabic, possibly other languages
  – Punctuation modeling for BN data
  – Incorporate machine learning algorithms
    • To reduce human annotation effort

• Guidelines, tools, progress, other details at
  www.ldc.upenn.edu/Projects/MDE