Reconstructing Spontaneous Speech

Erin Fitzgerald
Thesis Defense
January 19, 2009

Committee:
Frederick Jelinek and Sanjeev Khudanpur (Readers)
Andreas Andreou, Robert Frank, and Keith Hall
The difficulties of processing spontaneous speech

- Speakers make mistakes
  - Sentences not well-formed
  - Unpredictable structure
  - Automatic speech recognition (ASR) not always reliable
  - Information contained in acoustic inflections lost from transcripts

→ Problematic for extracting information from what was said

→ Extracted info needed for downstream applications
The difficulties of processing spontaneous speech

- yeah it's it's interesting that you say easter because i r-i remember as a a child i used to really really like easter.
- and i've i don't know i i guess as the kids were younger and growing up uh really liked it.
- an- and now easter has kind of.
- yeah.
- i i wouldn't say it's fallen into the background.
- but.
- yeah.
- it it doesn't seem to be quite as quite as big a holiday as it as it was uh years ago.
The difficulties of processing spontaneous speech

- yeah it's it's interesting that you say easter because i r- i remember as a as a child i used to really really like easter.
- and i've i don't know i i guess as the kids were younger and growing up uh really liked it.
- an- and now easter has kind of.
- yeah.
- i i wouldn't say it's fallen into the background.
- but.
- yeah.
- it it doesn't seem to be quite as quite as big a holiday as it as it was uh years ago.
The difficulties of processing spontaneous speech

- yeah it's it's interesting that you say easter because i remember as a child i used to really really like easter.
- and i've i don't know i guess as the kids were younger and growing up uh really liked it.
- an- and now easter has kind of.
- yeah.
- i i wouldn't say it's fallen into the background.
- but.
- yeah.
- it it doesn't seem to be quite as big a holiday as it was uh years ago.

Fillers: uh, um
The difficulties of processing spontaneous speech

- **yeah** it's interesting that you say Easter because I remember as a child I used to really really like Easter.
- **and** I've I don't know I guess as the kids were younger and growing up uh really liked it.
- **and** now Easter has kind of.
- **yeah.**
- **i** I wouldn't say it's really in the background.
- **but.**
- **yeah.**
- **it** it doesn't seem to be quite as quite as big a holiday as it was uh years ago.

**Fillers: uh, um**

**Contentless segments:** and, yeah
The difficulties of processing spontaneous speech

- yeah *it's* it's interesting that you say easter because *ir- i*
  remember *as a* as a child i used to *really really* like easter.
- and *i've i don't know* *i i guess* as the kids were younger
  and growing up uh *really liked it*.
- *an- and* now easter has kind of.
- *yeah*
- *i i wouldn't say* it's fallen into the background.
- *but.*
- *yeah.*
- *it* it doesn't seem to be *quite as* quite as big a holiday
  *as it* as it was uh *years ago*.

- **Fillers:** *uh, um*
- **Contentless segments:** *and, yeah*
- **Self repairs:** *it's it's, an- and*
The difficulties of processing spontaneous speech

- yeah it's it's interesting that you say easter because i r- i remember as a child i used to really really like easter.
- and i've i don't know i i guess as the kids were younger and growing up uh really liked it.
- an- and now easter has kind of.
- yeah.
- i i wouldn't say it's fallen into the background.
- but.
- yeah.
- it it doesn't seem to be quite as quite as big a holiday as it as it was uh years ago.

- Fillers: uh, um
- Contentless segments: and, yeah
- Self repairs: it's it's, an- and
- Parentheticals: I don't know, I guess as the kids...
The difficulties of processing spontaneous speech

- yeah it's it's interesting that you say easter because i r- i remember as a child i used to really really like easter.
- and i've i don't know i guess as the kids were younger and growing up i've really liked it.
- an- and now easter has kind of.
- yeah.
- i i wouldn't say it's fallen into the background.
- but.
- yeah.
- it it doesn't seem to be quite as quite as big a holiday as it as it was uh years ago.

- **Fillers:** uh, um
- **Contentless segments:** and, yeah
- **Self repairs:** it's it's, an- and
- **Parentheticals:** I don't know, I guess as the kids...
- **Reorderings:** I've + really liked it
The difficulties of processing spontaneous speech

- yeah it's it's interesting that you say easter because i r- i remember as a as a child i used to really really like easter.
- and i've i don't know i i guess as the kids were younger and growing up i've really liked it.
- an- and now easter has kind of <ARG>
- yeah.
- i i wouldn't say it's fallen into the background but.
- yeah.
- it it doesn't seem to be quite as quite as it as it was uh years ago.

- Fillers: uh, um
- Contentless segments: and, yeah
- Self repairs: it's it's, an- and
- Parentheticals: I don't know, I guess as the kids...
- Reorderings: I've + really liked it
- Missing arguments: <ARG>
The difficulties of processing spontaneous speech

- yeah it's interesting that you say easter because i remember as a child i used to really like easter.
- and i've don't know i guess as the kids were younger and growing up i've really liked it.
- an- and now easter has kind of.
- yeah.
- i wouldn't say it's fallen into the background.
- but.
- yeah.
- it doesn't seem to be quite as big a holiday.
- as it was years ago.

Fillers: uh, um
- Contentless segments: and, yeah
- Self repairs: it's, an- and
- Parentheticals: I don't know, I guess as the kids...
- Reorderings: I've + really liked it
- Missing arguments: <ARG>
What is Reconstruction?

The aim of speech reconstruction is to transform erroneous or disfluent spontaneous speech text into fluent and grammatical text that most closely expresses the meaning intended by the speaker.
Fluent and Grammatical

- **Fluency**: find and remove non-essential items
  - Filler words
  - Speaker revisions
  - Parenthetical asides
  - More

- **Grammaticality**: All sentence-like units (SU) containing verbs should have appropriate arguments (*and only those arguments*) in appropriate order and form
Minimal loss of content

- Sometimes revised data contains unique and preservable content

[but I think in general things in Europe] + I don’t think people are chain smokers necessarily

but I don’t think people in Europe are chain smokers necessarily
Motivation

- Downstream applications
  - Statistical Machine Translation (SMT): benefits shown from disfluency cleaning
  - Summarization
  - Question answering, etc.

- Transcript readability and comprehension

- ASR error rate is approaching human agreement
  
  *verbatim transcript isn’t what we really want!*

  - Work presented is transcript-based

---

Most common speaker errors: Simple Disfluencies

- Typically fall within the following types:

Reparandum Types

- Repetition:
  - [That’s] + {uh} that’s why I went there

- Revision:
  - But [when he] + {I mean} when she went there
  - It helps people [that are going to quit] + that would be quitting anyway

- Restart:
  - And [I think he’s] + he tells me he’s glad
  - [but I think in general things in Europe] + I don’t think people are chain smokers necessarily

Simple reconstruction requires only deletions

Reparandum Types

- Repetition:
  - [That’s] + {uh} that’s why I went there

- Revision:
  - But [when he] + {I mean} when she went there
  - It helps people [that are going to quit] + that would be quitting anyway

- Restart:
  - And [I think he’s] + he tells me he’s glad
  - [but I think in general things in Europe] + I don’t think people are chain smokers necessarily

Simple reconstruction requires only deletions

Simple Disfluencies Aren’t Enough

- In annotated reconstruction data, 20% of repaired sentences required more than a deletion.

We know that the Joe Millionaires are not going anywhere.

- Three reparanda (the$_1$, that$_2$, and ’s$_1$)
- Three word movements
- One deleted co-reference (that$_3$)
- One morphology substitution (’s$_2$ → are).
Goals of Work

I. Comprehensive understanding and unified analysis behind speaker-generated errors in spontaneous speech

II. Apply insights to statistical frameworks for automatic detection and correction of given phenomena.
Part I: Comprehensive error analysis

- Manual annotation effort
- Verify consistency of efforts
- Empirical data analysis
- Underlying (syntactic, semantic) analysis of changes made

Aim: Collect insights for both speech production models and feature development for automatic models
Part I: Comprehensive error analysis

- Manual annotation effort
  - Data selection and filtering
  - Reconstruction levels annotated
- Verify consistency of efforts
- Empirical data analysis
- Underlying (syntactic, semantic) analysis of changes made

Aim: Collect insights for both speech production models and feature development for automatic models
Manual annotation effort: Appropriate data selection

- Choose data to ensure relevance for continued research
- Fisher Conversational Telephone Speech:
  - Spontaneous (not read) speech pre-transcribed
  - Trees, simple disfluencies annotated for 21k SUs
  - Low ASR error rate (11.7% WER)
  - Ample examples of poor constructions
Identifying Poor Constructions:

1. Edit detection

- Apply edit detection
  - Noisy-channel “Tree Adjoining Grammar” (TAG) model targets simple edits

- Classifier:
  If one or more reparanda identified, label as “poor”

- Intuition: Overt repairs may be a good indicator of deeper errors

Identifying Poor Constructions:  
2. Plausible Deep-Syntax Parses

- Head-driven Phrase Structure Grammar (HPSG):
  - Deep-syntax, context-sensitive, hand-built grammar
  - More explanatory power than context-free parsers
  - No smoothing in implementation, so breaks for input not included in grammar

- Classifier:
  All sentences which produce no parse are “poor”

- Intuition: If deep structure is parseable,
  → good sentence

## Identifying Poor Constructions: Results

<table>
<thead>
<tr>
<th>Approach</th>
<th>“Poor” Classifier</th>
<th>P</th>
<th>R</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit detection</td>
<td>0/1: Edit found</td>
<td>96.0</td>
<td>73.5</td>
<td>83.3</td>
</tr>
<tr>
<td>HPSG</td>
<td>0/1: Not parseable</td>
<td>78.7</td>
<td>67.4</td>
<td>72.6</td>
</tr>
<tr>
<td>Edit U HPSG</td>
<td></td>
<td>80.4</td>
<td>87.5</td>
<td>83.8</td>
</tr>
</tbody>
</table>

- Approaches evaluated with precision, recall, and their weighted harmonic means on **500 SU random sample**.
- We extracted the union of these two approaches – **30% of the data** (6,384 of 21,456 SUs) – likely to contain errors to be annotated with reconstructions.
- Some bias here (certain error type missed?) which should be considered
Annotating Reconstructions: Levels of annotation

Includes options for

- Tracking alignments between original and reconstruction
  - Deleting (fillers, revisions, restarts, parentheticals, extraneous,...)
  - Inserting (function words, auxiliaries, null arguments)
  - Substituting (morphology, colloquial terms, transcript errors)
  - Reordering (arguments, adjuncts, other grammar)
- Adding and removing SU boundaries
Annotating Reconstructions: Levels of annotation

Includes options for

- Tracking alignments between original and reconstruction
  - Deleting (fillers, revisions, restarts, parentheticals, extraneous,...)
  - Inserting (function words, auxiliaries, null arguments)
  - Substituting (morphology, colloquial terms, transcript errors)
  - Reordering (arguments, adjuncts, other grammar)
- Adding and removing SU boundaries
Annotating Reconstructions: Levels of annotation

Includes options for
- Reconstruction efficacy labeled (five classes)
  - Cannot repair
  - Backchannel (“uh-huh”, “sure”)
  - Fragment, no content
  - Well-formed fragment
  - Well-formed and grammatical
- Grammatical reconstructions only: semantic roles labeled
Semantic Role Labeling (SRL): assign simple who did what to whom when, where, why, how... structure to sentence (PropBank role labels)

“We know that the Joe Millionaires are not going anywhere.”

Purpose:

- Compare to SRL structure for non-speech text
- SRL corpus for speech → automatic labeling, apps
- If automatic SRL labeling could be learned on speech, powerful reconstruction feature

Part I: Comprehensive error analysis

- Manual annotation effort
- Verify consistency of efforts
- Empirical data analysis
- Underlying (syntactic, semantic) analysis of changes made
Part I: Comprehensive error analysis

- Manual annotation effort
- Verify consistency of efforts
  - Is the data usable?
  - How difficult is the task?
- Empirical data analysis
  - Identify robust and statistically verifiable trends
- Underlying (syntactic, semantic) analysis of changes made
Annotation Characteristics

- 6,384 SUs labeled 2-3x each
- Average SU length drop ratio: 0.73
- Original vs. reconstructions:
  - minTER: 38.7% (i.e. if we did nothing)

\[
\min TER = \arg \min_{r \in R} \frac{\left| \{\text{dels, ins, subs, phrase moves}\} \right|}{\text{length of reference } R}
\]

→ Measure of changes made and baseline for evaluation comparison
Annotation Characteristics

● Breakdown of changes made:

● Breakdown of deletion types
Annotator (Pairwise) Agreement

- Only 52% of reconstructed strings match, but 92% of words match (annotator statistics)
- 84% of efficacy type labels match
- Avg inter-reconstruction minTER: 14.5% (best possible case)
- 83% of original words’ alignment labels match
  - Most common disagreement: revision vs. false start

Part I: Comprehensive error analysis

- Manual annotation effort
- Verify consistency of efforts
- Empirical data analysis
- Underlying (syntactic, semantic) analysis of changes made
Part I:
Comprehensive error analysis

- Manual annotation effort
- Verify consistency of efforts
- Empirical data analysis
- Underlying (syntactic, semantic) analysis of changes made
Underlying analysis of speaker-generated errors

- Semantic Role Labeling (SRL): propagate labels from reconstructions onto original string and analyze

- Further propagated onto automatic parses

- Details available, given time and interest
Part I Summary:
Comprehensive error analysis

- Produced richly annotated corpus of reconstructions
- Verified reconstruction quality
- Quantified reconstructive actions
- Examined the **structural and semantic differences** between spontaneous and cleaned speech
Goals of Work

I. Comprehensive understanding and unified analysis behind speaker-generated errors in spontaneous speech

II. Apply insights to statistical frameworks for automatic detection and correction of given phenomena.
Part II: Automatic identification and correction of errors

- Review previous and related efforts
- **System #1**: Word-level tagging and correction of simple errors
- **System #2**: SU-level error identification
- **System #3**: SU + word error identification and corrections
Part II: Automatic identification and correction of errors

- Review previous and related efforts
- System #1: Word-level tagging and correction of simple errors
- System #2: SU-level error identification
- System #3: SU + word error identification and corrections
Previous work

- 2004 NIST RT04 Metadata Extraction evaluation
- **Goal**: automatically identify fillers, reparanda, sentence-like unit (SU) boundaries
- **Results**: 

\[
\text{Edit Word Error Rate (EWER)} = \frac{|\text{false edits}| + |\text{missed edits}|}{|\text{true edits}|}
\]

<table>
<thead>
<tr>
<th>Data type</th>
<th>EWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR output</td>
<td>75%</td>
</tr>
<tr>
<td>ASR output + transcriptions</td>
<td>42%</td>
</tr>
</tbody>
</table>

Previous work: Acoustics

- **Prosody**:= the sound patterns of language, especially intonation, stress, and syllable timing
- Prosody is primary feature, but these systems outperformed by all-text system in NIST RT04 evaluation
  - More to gain still from text

Previous work: Tree Adjoining Grammars

- Johnson & Charniak 2004\(^1\) approach
- Uses *tree adjoining grammars* to capture *rough copy* edit regions

\[
\text{[the dog] \{I mean\} the cat ran quickly}
\]

- Rough copy (RC):= the same or similar words repeated in roughly the same order
- Zhang and Weng\(^2\) extended RC definition and showed improvements on SWBD edits

---

Reparandum Types

- Repetition:
  - [That’s] + {uh} that’s why I went there

- Revision:
  - But [when he] + {I mean} when she went there
  - It helps people [that are going to quit] + that would be quitting anyway

- Restart:
  - And [I think he’s] + he tells me he’s glad
  - [but I think in general things in Europe] + I don’t think people are chain smokers necessarily

Reparandum Types

- **Repetition:**
  - [That’s] + {uh} that’s why I went there

- **Revision:**
  - But [when he] + {I mean} when she went there
  - It helps people [that are going to quit] + that would be quitting anyway

- **Restart:**
  - And [I think he’s] + he tells me he’s glad
  - [but I think in general things in Europe] + I don’t think people are chain smokers necessarily

---

Reparandum Types

- **Repetition:**
  - [That’s] + {uh} that’s why I went there

- **Revision:**
  - But [when he] + {I mean} when she went there
  - It helps people [that are going to quit] + that would be quitting anyway

- **Restart:**
  - And [I think he’s] + he tells me he’s glad
  - [but I think in general things in Europe] + I don’t think people are chain smokers necessarily

Previous disfluency detection

<table>
<thead>
<tr>
<th>Label</th>
<th>% of words</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit (reparandum)</td>
<td>7.8%</td>
<td>85%</td>
<td>68%</td>
<td>76%</td>
</tr>
</tbody>
</table>

Data:
- **Training:**
  - Switchboard conversational telephone speech (CTS)
- **Testing:**
  - Spontaneous Speech Reconstruction (SSR) corpus (atop Fisher CTS corpus)
Previous disfluency detection

<table>
<thead>
<tr>
<th>Label</th>
<th>% of words</th>
<th>Precision</th>
<th>Recall</th>
<th>F₁-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit (reparandum)</td>
<td>7.8%</td>
<td>85%</td>
<td>68%</td>
<td>76%</td>
</tr>
</tbody>
</table>

E   E   E   E   E   FP   FL   -   -   -   -

did you the dog I mean the cat ran quickly

NC   NC   RC   RC   FL   FL   -   -   -   -

<table>
<thead>
<tr>
<th>Label</th>
<th>% of edits</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rough copy (RC) edits</td>
<td>58.8%</td>
<td>84.8%</td>
</tr>
<tr>
<td>Non-copy (NC) edits</td>
<td>41.2%</td>
<td>43.2%</td>
</tr>
<tr>
<td>Total edits</td>
<td>100.0%</td>
<td>67.6%</td>
</tr>
</tbody>
</table>
**Inspiration: Related NLP work**

- Deletion-based reconstruction → word tagging task

- Reconstruction → string transformation task

- Consider
  - Sentence **paraphrasing**
  - Document **summarization**
  - **Statistical machine translation**
    - (“bad” English to “good”)

Each have promise, but also limitations.
Part II: Automatic identification and correction of errors

- Review previous efforts
- **System #1:** Word-level tagging and correction of simple errors
- **System #2:** SU-level error identification
- **System #3:** SU + word error identification and cleaning
Conditional Random Fields for Word-level Error identification

- CRFs: undirected graphical models
  - Discriminative, not generative (like Maximum Entropy)
  - Model sequential context (like HMMs) *(globally normalized)*
- Widely applied for tagging tasks like POS tagging, sentence boundaries detection, etc.

\[
p_\lambda(Y|X) = \frac{1}{Z_\lambda(X)} \exp\left(\sum_k \lambda_k F_k(X, Y)\right)
\]

*Z_\lambda(X)* is a global normalization factor
\(\Lambda = (\lambda_1 \ldots \lambda_K)\) are model parameters related to each feature function \(F_k(X, Y)\)

\(Y\) represents the error class of each word
\(X\) represents observable inputs for each word

Lafferty, J.; McCallum, A. & Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data
Proceedings of the 18th International Conf. on Machine Learning, 2001, 282-289
Feature functions used

- **LEX:** Lexical features, including
  - the *lexical item + POS* for tokens $t_i$ and $t_{i+1}$
  - *distance* from previous token to the matching word/POS
  - Is previous token a *partial word*?
    What is distance to the next word with same start?
  - the token’s *position* within the sentence

- **JC04:** whether previous, next, or current word would be identified by the Johnson and Charniak (2004) system as an edit and/or a filler.

Feature functions used

- **LM**: Language model features to capture juxtaposition of the reparandum and the repair
  - (unigram) $\log p(t_{i+1})$ for the next word/POS token $t_{i+1}$
  - (conditional) $\log p(t_{i+1}|h)$ given multi-token history $h$
  - Ratio $\log [p(t|h)/ p(t)]$: approximation for $I(t;h)$ (derived)

- Non-terminal (**NT**) ancestors: (on automatic parse)
  For each word, we identified
  - NT phrases just completed (if any)
  - NT phrases about to start to its right (if any)
  - NT constituents for which the word is included.

Useful for NCs
Experimental Designs (3 param)

- Train: full or errorful-only data
- Test: full or errorful-only data
- Classes: Identify all vs. some classes

<table>
<thead>
<tr>
<th></th>
<th>Train data</th>
<th>Test data</th>
<th>Classes trained by model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Full train</td>
<td>Full test</td>
<td>FL + RC + NC</td>
</tr>
<tr>
<td>2</td>
<td>Full train</td>
<td>Full test</td>
<td>FL + RC + NC, FL+{RC, NC}</td>
</tr>
<tr>
<td>3</td>
<td>Errorful SUs</td>
<td>Errorful SUs</td>
<td>FL + RC + NC</td>
</tr>
<tr>
<td>4</td>
<td>Errorful SUs</td>
<td>Full test</td>
<td>FL + RC + NC</td>
</tr>
<tr>
<td>5</td>
<td>Errorful SUs</td>
<td>Errorful SUs</td>
<td>{RC, NC}, FL+{RC, NC}</td>
</tr>
<tr>
<td>6</td>
<td>Errorful SUs</td>
<td>Full test</td>
<td>{RC, NC}, FL+{RC, NC}</td>
</tr>
</tbody>
</table>
Evaluation Metric #1: Label P/R/F using two references

<table>
<thead>
<tr>
<th>Setup</th>
<th>Class labeled</th>
<th>Training</th>
<th>Testing</th>
<th>FL</th>
<th>RC</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>FP+RC+NC</td>
<td>All train</td>
<td>All test</td>
<td>71.0</td>
<td>80.3</td>
<td>47.4</td>
</tr>
<tr>
<td>#2</td>
<td>NC</td>
<td>All train</td>
<td>All test</td>
<td>-</td>
<td>80.3</td>
<td>42.5</td>
</tr>
<tr>
<td>#2</td>
<td>NC+FL</td>
<td>All train</td>
<td>All test</td>
<td>70.8</td>
<td>-</td>
<td>47.5</td>
</tr>
<tr>
<td>#2</td>
<td>RC</td>
<td>All train</td>
<td>All test</td>
<td>70.8</td>
<td>84.2</td>
<td>-</td>
</tr>
<tr>
<td>#2</td>
<td>RC+FL</td>
<td>All train</td>
<td>All test</td>
<td>67.8</td>
<td>84.7</td>
<td>-</td>
</tr>
</tbody>
</table>

- Training for individual features yields some improvement over joint training
- Given accurate poor sentence ID, sizable improvements possible.
Evaluation Metric #2:
Feature class comparison of P/R/F

- RC identification better than NC
- NT features useful for both RC/NC
- LM features not powerful on their own
- NC unexpectedly LEX-sensitive
- NCs still share some lexical items and structure
- Partial word feature very important (but more realistic to delete)
- RC very resilient (repeated info across features)

<table>
<thead>
<tr>
<th>Features</th>
<th>FL</th>
<th>RC</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>JC04 only</td>
<td>56.6</td>
<td>69.9-81.9</td>
<td>1.6-21.0</td>
</tr>
<tr>
<td>lexical only</td>
<td>56.5</td>
<td>72.7</td>
<td>33.4</td>
</tr>
<tr>
<td>LM only</td>
<td>0.0</td>
<td>15.0</td>
<td>0.0</td>
</tr>
<tr>
<td>NT bounds only</td>
<td>44.1</td>
<td>35.9</td>
<td>11.5</td>
</tr>
<tr>
<td>All but JC04</td>
<td>58.5</td>
<td>79.3</td>
<td>33.1</td>
</tr>
<tr>
<td>All but lexical</td>
<td>66.9</td>
<td>76.0</td>
<td>19.6</td>
</tr>
<tr>
<td>All but LM</td>
<td>67.9</td>
<td>83.1</td>
<td>41.0</td>
</tr>
<tr>
<td>All but NT bounds</td>
<td>61.8</td>
<td>79.4</td>
<td>33.6</td>
</tr>
<tr>
<td>All</td>
<td>71.0</td>
<td>80.3</td>
<td>47.4</td>
</tr>
</tbody>
</table>
Evaluation Metric #3:  
Exact SU match (Reconstruction)

- Results summary:
  - Minimal deletions: 78% of test SUs match reference
  - CRF method: 2.3% improvement over JC04

- Considering only SUs with deeper errors...
  - Minimal deletions: 2% of SUs match
  - CRF: 56% match (11% improvement over JC04)
Evaluation Metric #3: Exact SU match (Reconstruction)

\[ SU_{\text{match}} = \frac{1}{S} \sum_{s \in S} \max_{r \in R} \delta(s, r) \]

<table>
<thead>
<tr>
<th>Setup</th>
<th>Classes deleted</th>
<th># SUs</th>
<th># SUs which match gold</th>
<th>% accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>only filled pauses</td>
<td>2288</td>
<td>1800</td>
<td>78.7%</td>
</tr>
<tr>
<td>JC04-1</td>
<td>E+FL</td>
<td>2288</td>
<td>1858</td>
<td>81.2%</td>
</tr>
<tr>
<td>CRF-#1</td>
<td>RC, NC, and FL</td>
<td>2288</td>
<td>1922</td>
<td>84.0%</td>
</tr>
<tr>
<td>CRF-#2</td>
<td>( \bigcup {RC,NC} )</td>
<td>2288</td>
<td>1901</td>
<td>83.1%</td>
</tr>
</tbody>
</table>

Accuracy appears much higher when non-errorful utterances included

- Our method shows improvement over JC04
  - Even more if we could select errorful SUs first! (improves precision)
Part II: Automatic identification and correction of errors

- Review previous and related efforts
- **System #1:** Word-level tagging and correction of simple errors
- **System #2:** SU-level error identification
- **System #3:** SU + word error identification and corrections
Identifying Poor Constructions

- To capture interaction between features, we use maximum entropy classification system
  - Like CRFs, discriminative classifier

- Overview of features

- Overview of feature combinations

- Results (on manually labeled SSR test set)
SU-level error ID features: 
JC04 detection and HPSG parsing

- Applied together for pre-filtering of manually annotated data
  - SUs with one or more reparandum word labeled by JC04
  - SUs left unparsed by deep syntax HPSG grammar

→ Marked by binary indicator functions

SU-level error ID features: Length and Backchannels

- **Intuition #1**: Long SUs tend to have more errors; short SUs are often backchannels.

- **Intuition #2**: Reconstruction unnecessary for backchannel acknowledgements:
  - “uh-huh”, “sure”, etc.
  and likely same for combinations.
SU-level error ID features:
Unseen auto parse constructions

- **Intuition:** Robust parsers may generate new rules when faced with unfamiliar context (e.g. Charniak Markov grammar)

SU-level error ID features:

Unseen auto parse C-commands

- Node $A$ c-commands node $B$ if
  - Neither $A$ nor $B$ dominate each other
  - $A$ immediately precedes $B$ at first branching node

- Intuition for inclusion:
  Parsers condition on dominance and sisters; 
  left c-command shows longer (richer NT) missed by parser
SU-level error ID features:
Unseen auto parse C-commands

Rule expansions + c-commanding NT:

<table>
<thead>
<tr>
<th>Rule</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow NP \ VP$</td>
<td></td>
</tr>
<tr>
<td>$NP_1 \rightarrow PRP$</td>
<td></td>
</tr>
<tr>
<td>$VP_1 \rightarrow V \ SBAR$</td>
<td></td>
</tr>
<tr>
<td>$SBAR \rightarrow S$</td>
<td></td>
</tr>
<tr>
<td>$S \rightarrow NP_2 \ VP_2$</td>
<td>VBP</td>
</tr>
<tr>
<td>$NP_2 \rightarrow DT$</td>
<td></td>
</tr>
<tr>
<td>$VP_2 \rightarrow VBZ \ VP$</td>
<td></td>
</tr>
<tr>
<td>$VP_3 \rightarrow VBG \ NP$</td>
<td></td>
</tr>
<tr>
<td>$NP_3 \rightarrow DT \ NN$</td>
<td></td>
</tr>
</tbody>
</table>

*no left c-command*
SU-level error ID features:

Unseen auto parse C-commands

- Node $A$ **c-commands** node $B$ if
  - Neither $A$ nor $B$ dominate each other
  - $A$ immediately precedes $B$ at first branching node

- **Intuition for inclusion:**
  Parsers condition on dominance and sisters;
  **left c-command** shows longer (richer NT) missed by parser
## Feature combination results

<table>
<thead>
<tr>
<th>Features included</th>
<th>JC04</th>
<th>HPSG</th>
<th>Rules</th>
<th>C-comm</th>
<th>Length</th>
<th>Backchannel</th>
<th>F$_1$-score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>a) Individual features</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>79.9</td>
</tr>
<tr>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>77.1</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>59.7</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>42.2</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>23.2</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>√</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>b) All features combined</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>83.3</td>
</tr>
<tr>
<td><strong>c) All-but-one</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>78.4 (-4.9)</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>81.2 (-2.1)</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>81.3 (-2.0)</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>82.1 (-1.2)</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>82.9 (-0.4)</td>
</tr>
<tr>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>X</td>
<td>√</td>
<td>83.2 (-0.1)</td>
</tr>
</tbody>
</table>
Part II: Automatic identification and correction of errors

- Review previous and related efforts
- **System #1**: Word-level tagging and correction of simple errors
- **System #2**: SU-level error identification
- **System #3**: SU + word error identification and corrections
Combined model: word-level correction on auto ID’d SUs

**NC prediction**

**RC prediction**

- **All training data, all testing data**
- **Train with reference errors, test with auto ID’d errors**
- **Train with reference errors, test with ref errors**
Combined model: SU match for FL/RC/NC reconstruction

<table>
<thead>
<tr>
<th>Setup</th>
<th>Classed deleted</th>
<th>Testing</th>
<th># SUs (filt/unfilt)</th>
<th># SUs that match ref</th>
<th>% accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-1</td>
<td>only filled pauses</td>
<td>All data</td>
<td>2288</td>
<td>1800</td>
<td>78.7%</td>
</tr>
<tr>
<td>JC04-1</td>
<td>E+FL</td>
<td>All data</td>
<td>2288</td>
<td>1858</td>
<td>81.2%</td>
</tr>
<tr>
<td>MaxEnt/CRF-1</td>
<td>FL+RC+NC</td>
<td>All data</td>
<td>2288</td>
<td>1922</td>
<td>84.0%</td>
</tr>
<tr>
<td>Baseline-2</td>
<td>only filled pauses</td>
<td>Auto ID’d</td>
<td>430</td>
<td>84</td>
<td>19.5%</td>
</tr>
<tr>
<td>JC04-2</td>
<td>E+FL</td>
<td>Auto ID’d</td>
<td>430</td>
<td>187</td>
<td>43.5%</td>
</tr>
<tr>
<td>MaxEnt/CRF-2</td>
<td>FL+RC+NC</td>
<td>Auto ID’d</td>
<td>430</td>
<td>223</td>
<td>51.9%</td>
</tr>
<tr>
<td>Baseline-3</td>
<td>only filled pauses</td>
<td>Gold errors</td>
<td>281</td>
<td>5</td>
<td>1.8%</td>
</tr>
<tr>
<td>JC04-3</td>
<td>E+FL</td>
<td>Gold errors</td>
<td>281</td>
<td>126</td>
<td>44.8%</td>
</tr>
<tr>
<td>MaxEnt/CRF-3</td>
<td>FL+RC+NC</td>
<td>Gold errors</td>
<td>281</td>
<td>156</td>
<td>55.5%</td>
</tr>
</tbody>
</table>
Part II Summary: Automatic Reconstruction

- Used CRFs for word level error class labeling
  - Analyzed impact of orthogonal features
  - Varied experimental setup
  - Two evaluation metrics
- SU-level errors identified via maxent model

- Source of gains: richer data, and parameters tuned for more specific error types
Summary of Contributions and Conclusions

- Motivated and presented a framework for the reconstruction of spontaneous speech
- Produced richly annotated corpus of reconstructions, providing training, evaluation, and motivated feature design via insights
- Implemented reconstruction model to better identify restart fragments and other reparanda
- Presented methods for identifying poorly constructed sentences

- Work and analysis should motivate further speech reconstruction work in years ahead!
Acknowledgements

- NSF PIRE support
- Fred Jelinek and thesis committee
- Johns Hopkins: Carolina Parada
- Charles University: Petr Podveský, Petr Pajas, Silvie Cinková
- Saarland University: Yi Zhang and Valia Kordoni
- Many useful conversations with many helpful colleagues!
Publications