Outline

- Advanced in Search
  - Multipass search
    - N-best lists
    - Lattices
- Advances in Context
  - Context-dependent (triphone) acoustic modeling
- Metadata
  - Disfluencies, etc
yeah actually um i belong to a gym down here a gold’s gym uh-huh and uh exercise i try to exercise five days a week um and i usually do that uh what type of exercising do you do in the gym

A: Yeah I belong to a gym down here. Gold’s Gym. And I try to exercise five days a week. And I usually do that.

B: What type of exercising do you do in the gym?
Metadata tasks

- Diarization
- Sentence Boundary Detection
- Truecasing
- Punctuation detection
- Disfluency detection
Why?

- Diarization?
- Sentence Boundary Detection?
Sentence Segmentation (Shriberg et al 2000)

- Binary classification task; judge the juncture between each two words:

- Features:
  - Pause
  - Duration of previous phone and rime
  - Pitch change across boundary; pitch range of previous word

---

without drinking water due to the flood (pause) many communities are still cut off...
Disfluencies: standard terminology (Levelt)

- **Reparandum**: thing repaired
- **Interruption point (IP)**: where speaker breaks off
- **Editing phase (edit terms)**: uh, I mean, you know
- **Repair**: fluent continuation
# Types of Disfluencies

<table>
<thead>
<tr>
<th>Disfluency type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>fillers (or filled pauses):</td>
<td>But, <em>uh</em>, that was absurd</td>
</tr>
<tr>
<td>word fragments</td>
<td>A guy went to a <em>d</em>- , a landfill</td>
</tr>
<tr>
<td>repetitions:</td>
<td>it was just a <em>change of, change of</em> location</td>
</tr>
<tr>
<td>restarts</td>
<td>it’s – I find it very strange</td>
</tr>
</tbody>
</table>
Why Disfluencies?

- Cues to emotion or attitude
- Known to cause problems for speech recognition
  - Goldwater, Jurafsky, Manning (2008):
    - Word errors *increase by up to 15% (absolute)* for words near fragments
Why disfluencies?

- Need to clean them up to get understanding
  - Does American airlines offer any one-way flights [uh] one-way fares for 160 dollars?
  - Delta leaving Boston seventeen twenty one arriving Fort Worth twenty two twenty one forty

- Known to cause problems for speech recognition
  - Goldwater, Jurafsky, Manning (2008):
    - Word errors increase by up to 15% (absolute) for words near fragments

- Might help in language modeling
  - Disfluencies might occur at particular positions (boundaries of clauses, phrases)

- Annotating them helps readability
Counts (from Shriberg, Heeman)

- **Sentence disfluency rate**
  - ATIS: 6% of sentences disfluent (10% long sentences)
  - Levelt human dialogs: 34% of sentences disfluent
  - Swbd: ~50% of multiword sentences disfluent
  - TRAINS: 10% of words are in reparandum or editing phrase

- **Word disfluency rate**
  - SWBD: 6%
  - ATIS: 0.4%
  - AMEX 13%
    - (human-human air travel)
Prosodic characteristics of disfluencies

- Nakatani and Hirschberg 1994
- Fragments are good cues to disfluencies
- Prosody:
  - Pause duration is shorter in disfluent silence than fluent silence
  - F0 increases from end of reparandum to beginning of repair, but only minor change
  - Repair interval offsets have minor prosodic phrase boundary, even in middle of NP:
    - Show me all n- | round-trip flights | from Pittsburgh | to Atlanta
Syntactic Characteristics of Disfluencies

- Hindle (1983)
- The repair often has same structure as reparandum
- Both are Noun Phrases (NPs) in this example:

```
Does United offer any one-way flights uh, I mean, one-way fares for 160 dollars?
```

- So if could automatically find IP, could find and correct reparandum!
Disfluencies and LM

- Clark and Fox Tree
- Looked at “um” and “uh”
  - “uh” includes “er” (“er” is just British non-rhotic dialect spelling for “uh”)
- Different meanings
  - Uh: used to announce minor delays
    - Preceded and followed by shorter pauses
  - Um: used to announce major delays
    - Preceded and followed by longer pauses
Um versus uh: delays (Clark and Fox Tree)

Fig. 1. Percent of fillers followed by delays (LL corpus).

Fig. 2. Mean length of pauses after fillers (LL corpus).
The more difficulty speakers have in planning, the more delays.

Consider 3 locations:

- I: before intonation phrase: hardest
- II: after first word of intonation phrase: easier
- III: later: easiest

And then uh somebody said, . [I] but um -- [II] don’t you think there’s evidence of this, in the twelfth - [III] and thirteenth centuries?
Fragments

- Incomplete or cut-off words:
  - Leaving at seven fif- eight thirty
  - uh, I, I d-, don't feel comfortable
  - You know the fam-, well, the families
  - I need to know, uh, how- how do you feel...

- Uh yeah, yeah, well, it- it- that’s right. And it-

- SWBD: around 0.7% of words are fragments (Liu 2003)
- ATIS: 60.2% of repairs contain fragments (6% of corpus sentences had a least 1 repair) Bear et al (1992)
- Another ATIS corpus: 74% of all reparanda end in word fragments (Nakatani and Hirschberg 1994)
Fragment glottalization

- Uh yeah, yeah, well, **it- it-** that’s right. And **it-**
Fragments in other languages

- Mandarin (Chu, Sung, Zhao, Jurafsky 2006)
- Fragments cause similar errors as in English:

Substitution: 你 - 你 下次 跟他 说

\textit{you-you next time to him tell}

Recognizer output: 那 你 下次 跟他 说

\textit{that you next time to him tell}

- 但我 - 我问的是
- I - I was asking...
- 他是 - 却很 - 活的很好
- He very - lived very well
Fragments in Mandarin

- Mandarin fragments unlike English; no glottalization.
- Instead: Mostly (unglottalized) repetitions

a. 我 - 我问的是 有什么影响
   I - I ask DE copula have what influence
   ‘I asked what influence it has.’

b. 他 却 很 - 活的 很好
   he but very - live very well
   ‘But he lives very well.’

- So: best features are lexical, rather than voice quality
Does deleting disfluencies improve LM perplexity?

- Stolcke and Shriberg (1996)
- Build two LMs
  - Raw
  - Removing disfluencies
- See which has a higher perplexity on the data.
  - Filled Pause
  - Repetition
  - Deletion
Change in Perplexity when Filled Pauses (FP) are removed

- LM Perplexity goes up at following word:

<table>
<thead>
<tr>
<th>Position</th>
<th>ĈH</th>
<th>ĈH+1</th>
<th>ĈH+2</th>
<th>ĈM</th>
<th>ĈM+1</th>
<th>ĈM+2</th>
<th>non-FP</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>39.0</td>
<td>223.5</td>
<td>89.8</td>
<td>174.9</td>
<td>36.7</td>
<td>71.9</td>
<td>103.4</td>
<td>101.9</td>
</tr>
<tr>
<td>FP model</td>
<td>39.9</td>
<td>291.5</td>
<td>91.4</td>
<td>175.8</td>
<td>73.4</td>
<td>69.2</td>
<td>103.4</td>
<td>103.3</td>
</tr>
<tr>
<td>#events</td>
<td>502</td>
<td>502</td>
<td>373</td>
<td>188</td>
<td>188</td>
<td>94</td>
<td>19426</td>
<td></td>
</tr>
</tbody>
</table>

- Removing filled pauses makes LM worse!!
- I.e., filled pauses seem to help to predict next word.
- Why would that be?
Filled pauses tend to occur at clause boundaries

- Word before FP is end of previous clause; word after is start of new clause;
  - Best not to delete FP
- Some of the different things we’re doing there’s not time to do it all
- “there’s” is very likely to start a sentence
- So $P(\text{there’s}|\text{uh})$ is better estimate than $P(\text{there’s}|\text{doing})$
Suppose we just delete medial FPs

- **Experiment 2:**
  - Parts of SWBD hand-annotated for clauses
  - Build FP-model by deleting only medial FPs
  - Now prediction of post-FP word (perplexity) improves greatly!
  - Siu and Ostendorf found same with “you know”

<table>
<thead>
<tr>
<th>Position</th>
<th>CH+1</th>
<th>CM+1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>849.0</td>
<td>437.4</td>
</tr>
<tr>
<td>FP model</td>
<td>606.2</td>
<td>361.7</td>
</tr>
</tbody>
</table>
What about REP and DEL

- S+S built a model with “cleaned-up” REP and DEL
- Slightly lower perplexity
- But exact same word error rate (49.5%)
- Why?
  - Rare: only 2 words per 100
  - Doesn’t help because adjacent words are misrecognized anyhow!
Stolcke and Shriberg conclusions wrt LM and disfluencies

- Disfluency modeling purely in the LM probably won’t vastly improve WER
- But
  - Disfluencies should be considered jointly with sentence segmentation task
  - Need to do better at recognizing disfluent words themselves
  - Need acoustic and prosodic features
WER reductions from modeling disfluencies + background events

- Rose and Riccardi (1999)

**Figure 1:** a) Inclusion of all labeled background events (LBEs) in a single “between-word” loop. b) Portion of phrase-based LM trained from LBE annotated text.
HMIHY Background events

- Out of 16,159 utterances:

<table>
<thead>
<tr>
<th>Event</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filled Pauses</td>
<td>7189</td>
</tr>
<tr>
<td>Word Fragments</td>
<td>1265</td>
</tr>
<tr>
<td>Hesitations</td>
<td>792</td>
</tr>
<tr>
<td>Laughter</td>
<td>163</td>
</tr>
<tr>
<td>Lipsmack</td>
<td>2171</td>
</tr>
<tr>
<td>Breath</td>
<td>8048</td>
</tr>
<tr>
<td>Non-Speech Noise</td>
<td>8834</td>
</tr>
<tr>
<td>Background Speech</td>
<td>3585</td>
</tr>
<tr>
<td>Operator Utt.</td>
<td>5112</td>
</tr>
<tr>
<td>Echoed Prompt</td>
<td>5353</td>
</tr>
</tbody>
</table>
Phrase-based LM

- “I would like to make a collect call”
- “a [wfrag]”
- <dig3> [brth] <dig3>
- “[brth] and I”
Modeling “LBEs” does help in WER

<table>
<thead>
<tr>
<th>ASR Word Accuracy</th>
<th>Test Corpora</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System Configuration</strong></td>
<td><strong>HMM</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>Baseline</td>
</tr>
<tr>
<td>LBE</td>
<td>Baseline</td>
</tr>
<tr>
<td>LBE</td>
<td>LBE</td>
</tr>
</tbody>
</table>
More on location of FPs

- **Peters: Medical dictation task**
  - Monologue rather than dialogue
  - In this data, FPs occurred INSIDE clauses
  - Trigram PP after FP: 367
  - Trigram PP after word: 51

- **Stolcke and Shriberg (1996b)**
  - \( w_k \) FP \( w_{k+1} \): looked at \( P(w_{k+1} | w_k) \)
  - Transition probabilities lower for these transitions than normal ones

- **Conclusion:**
  - People use FPs when they are planning difficult things, so following words likely to be unexpected/rare/difficult
Detection of disfluencies

- Nakatani and Hirschberg
- Decision tree at $w_i$-$w_j$ boundary
  - pause duration
  - Word fragments
  - Filled pause
  - Energy peak within $w_i$
  - Amplitude difference between $w_i$ and $w_j$
  - F0 of $w_i$
  - F0 differences
  - Whether $w_i$ accented

- Results:
  - 78% recall/89.2% precision
Detection/Correction

- Bear, Dowding, Shriberg (1992)
- System 1:
  - Hand-written pattern matching rules to find repairs
    - Look for identical sequences of words
    - Look for syntactic anomalies (“a the”, “to from”)
    - 62% precision, 76% recall
    - Rate of accurately correcting: 57%
Using Natural Language Constraints

- Gemini natural language system
- Based on Core Language Engine
- Full syntax and semantics for ATIS
- Coverage of whole corpus:
  - 70% syntax
  - 50% semantics
Using Natural Language Constraints

- Gemini natural language system
- Run pattern matcher
- For each sentence it returned
  - Remove fragment sentences
  - Leaving 179 repairs, 176 false positives
  - Parse each sentence
    - If succeed: mark as false positive
    - If fail:
      - run pattern matcher, make corrections
      - Parse again
      - If succeeds, mark as repair
      - If fails, mark no opinion
NL Constraints

- Syntax Only
  - Precision: 96%
- Syntax and Semantics
  - Correction: 70%
Recent work: EARS Metadata Evaluation (MDE)

- A recent multiyear DARPA bakeoff
- Sentence-like Unit (SU) detection:
  - find end points of SU
  - Detect subtype (question, statement, backchannel)
- Edit word detection:
  - Find all words in reparandum (words that will be removed)
- Filler word detection
  - Filled pauses (uh, um)
  - Discourse markers (you know, like, so)
  - Editing terms (I mean)
- Interruption point detection

Liu et al 2003
Kinds of disfluencies

- Repetitions
  - I * I like it

- Revisions
  - We * I like it

- Restarts (false starts)
  - It’s also * I like it
Conventions:
- ./ for statement SU boundaries,
- <> for fillers,
- [ ] for edit words,
- * for IP (interruption point) inside edits

And <uh> <you know> wash your clothes wherever you are ./ and [ you ] * you really get used to the outdoors ./
## MDE Labeled Corpora

<table>
<thead>
<tr>
<th></th>
<th>CTS</th>
<th>BN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set (words)</td>
<td>484K</td>
<td>182K</td>
</tr>
<tr>
<td>Test set (words)</td>
<td>35K</td>
<td>45K</td>
</tr>
<tr>
<td>STT WER (%)</td>
<td>14.9</td>
<td>11.7</td>
</tr>
<tr>
<td>SU %</td>
<td>13.6</td>
<td>8.1</td>
</tr>
<tr>
<td>Edit word %</td>
<td>7.4</td>
<td>1.8</td>
</tr>
<tr>
<td>Filler word %</td>
<td>6.8</td>
<td>1.8</td>
</tr>
</tbody>
</table>
MDE Algorithms

- Use both text and prosodic features
- At each interword boundary
  - Extract Prosodic features (pause length, durations, pitch contours, energy contours)
  - Use N-gram Language model
  - Combine via HMM, Maxent, CRF, or other classifier
State of the art: SU detection

- 2 stage
  - Decision tree plus N-gram LM to decide boundary
  - Second maxent classifier to decide subtype

- Current error rates:
  - Finding boundaries
    - 40-60% using ASR
    - 26-47% using transcripts
State of the art: Edit word detection

- Multi-stage model
  - HMM combining LM and decision tree finds IP
  - Heuristics rules find onset of reparandum
  - Separate repetition detector for repeated words

- One-stage model
  - CRF jointly finds edit region and IP
  - BIO tagging (each word has tag whether is beginning of edit, inside edit, outside edit)

- Error rates:
  - 43-50% using transcripts
  - 80-90% using ASR
Using only lexical cues

- 3-way classification for each word
  - Edit, filler, fluent

- Using TBL
  - Templates: Change
    - Word X from L1 to L2
    - Word sequence X Y to L1
    - Left side of simple repeat to L1
    - Word with POS X from L1 to L2 if followed by word with POS Y
Rules learned

- Label all fluent filled pauses as fillers
- Label the left side of a simple repeat as an edit
- Label “you know” as fillers
- Label fluent well’s as filler
- Label fluent fragments as edits
- Label “I mean” as a filler
Error rates using only lexical cues

- **CTS, using transcripts**
  - Edits: 68%
  - Fillers: 18.1%

- **Broadcast News, using transcripts**
  - Edits 45%
  - Fillers 6.5%

- **Using speech:**
  - Broadcast news filler detection from 6.5% error to 57.2%

- **Other systems (using prosody) better on CTS, not on Broadcast News**
Conclusions: Lexical Cues Only

- Can do pretty well with only words
  - (As long as the words are correct)
- Much harder to do fillers and fragments from ASR output, since recognition of these is bad
Accents: An experiment

- A word by itself

- The word in context
Summary

- **Advanced in Search**
  - Multipass search
    - N-best lists
    - Lattices

- **Advances in Context**
  - Context-dependent (triphone) acoustic modeling

- **Metadata**
  - Disfluencies, etc