

Recognizing Metadata: Segmentation and Disfluencies

Julia Hirschberg

CS 4706

DARPA EARS Program

- Effective, Affordable, Reusable Speech-to-Text
- Goals: Produce transcriptions that are more readable and usable by humans and downstream processes
 - Segment speech into text-like units
 - Sentences, speaker turns, topics
 - Add appropriate punctuation
 - Eliminate spontaneous speech phenomena
 - Filled pauses, Self-repairs, Discourse markers
- Tasks: ASR, Metadata Extraction (MDE)

Motivation: ASR Transcription

- aides tonight in boston in depth the truth squad for special series until election day tonight the truth about the budget surplus of the candidates are promising the two international flash points getting worse while the middle east and a new power play by milosevic and a lifelong a family tries to say one child life by having another amazing breakthrough the u s was was told local own boss good evening uh from the university of massachusetts in boston the site of the widely anticipated first of eight between vice president al gore and governor george w bush with the election now just five weeks away this is the beginning of a sprint to the finish and a strong start here tonight is important this is the stage for the two candidates will appear before a national television audience taking questions from jim lehrer of p b s n b c's david gregory is here with governor bush claire shipman is covering the vice president claire you begin tonight please

Motivation: Speaker Segmentation (Diarization)

- Speaker: 0 - aides tonight in boston in depth the truth squad for special series until election day tonight the truth about the budget surplus of the candidates are promising the two international flash points getting worse while the middle east and a new power play by milosevic and a lifelong a family tries to say one child life by having another amazing breakthrough the u s was was told local own boss good evening uh from the university of massachusetts in boston
- Speaker: 1 - the site of the widely anticipated first of eight between vice president al gore and governor george w bush with the election now just five weeks away this is the beginning of a sprint to the finish and a strong start here tonight is important this is the stage for the two candidates will appear before a national television audience taking questions from jim lehrer of p b s n b c's david gregory is here with governor bush claire shipman is covering the vice president claire you begin tonight please

Motivation: Sentence Detection, Punctuation, Truecasing

- Speaker: Anchor - Aides tonight in Boston. In depth the truth squad for special series until election day. Tonight the truth about the budget surplus of the candidates are promising. The two international flash points getting worse. While the Middle East. And a new power play by Milosevic and a lifelong a family tries to say one child life by having another amazing breakthrough the U. S. was was told local own boss. Good evening uh from the university of Massachusetts in Boston.
- Speaker: Reporter - The site of the widely anticipated first of eight between Vice President Al Gore and Governor George W. Bush. With the election now just five weeks away. This is the beginning of a sprint to the finish. And a strong start here tonight is important. This is the stage for the two candidates will appear before a national television audience taking questions from Jim Lehrer of PBS. NBC's David Gregory is here with Governor Bush. Claire Shipman is covering the vice president. Claire, you begin tonight please.

Story Boundary/Topic Detection

Speaker: Anchor - Aides tonight in Boston. In depth the truth squad for special series until election day. Tonight the truth about the budget surplus of the candidates are promising. The two international flash points getting worse. While the Middle East. And a new power play by Milosevic and a lifelong a family tries to say one child life by having another amazing breakthrough the U. S. was was told local own boss.

Good evening uh from the university of Massachusetts in Boston.

Speaker: Reporter - The site of the widely anticipated first of eight between Vice President Al Gore and Governor George W. Bush. With the election now just five weeks away. This is the beginning of a sprint to the finish. And a strong start here tonight is important. This is the stage for the two candidates will appear before a national television audience taking questions from Jim Lehrer of PBS.

NBC's David Gregory is here with Governor Bush. Claire Shipman is covering the vice president. Claire, you begin tonight please.

Today

- Segmentation
 - Speakers
 - Sentences
 - Stories
- Disfluency detection/correction
 - Self-Repairs

Speaker Diarization

- Assign consistent speaker labels across a meeting or news broadcast
 - Speaker1...
 - Speaker2...
 - Speaker1...
 - Speaker4...
- Segment spoken document into acoustically distinct units
- Cluster and assign identifiers to each instance in the document

Sentence Segmentation

- Classification task: sentence boundary vs. no sentence boundary
- Features:
 - Lexical and POS information (but ASR is noisy)
 - Distance from previous hypothesized boundary
 - Speech information
 - Durations (sentence-final words are longer)
 - Pause
 - F0 (f0 modeling, pitch reset, pitch range)
- Hand-annotated training corpus, annotated for SLUs (Sentence-like Units): 51pp [LDC Manual](#)

Punctuation Detection and Truecasing

- Punctuation:
 - Assign each SLU an appropriate final punctuation
 - SLU-internal punctuation?
- Capitalization:
 - Capitalize words beginning sentences
 - Named Entities – how train?
- Features:
 - Prosodic and lexical
- Training Data?

Topic/Story Boundary Detection

- Rich text-based literature
 - Halliday & Hasan 1976: lexical cohesion
 - Hearst 1997: TextTiling segments by comparing words before and after each hypothesized topic boundary wrt a word similarity metric
 - Reynar, 1999; Beeferman et al 1999: cue phrases
 - Choi 2000: divisive clustering using cosine sim on stems
- Features used:
 - Stem repetition, entity repetition, word frequency, context vectors, semantic similarity, word distance, lexical chains, anaphoric chains

Spoken Cues to Discourse/Topic Structure

- Pitch range

Lehiste '75, Brown et al '83, Silverman '86, Avesani & Vayra '88, Ayers '92, Swerts et al '92, Grosz & Hirschberg'92, Swerts & Ostendorf '95, Hirschberg & Nakatani '96

- Preceding pause

Lehiste '79, Chafe '80, Brown et al '83, Silverman '86, Woodbury '87, Avesani & Vayra '88, Grosz & Hirschberg'92, Passoneau & Litman '93, Hirschberg & Nakatani '96

- Rate

Butterworth '75, Lehiste '80, Grosz & Hirschberg'92,
Hirschberg & Nakatani '96

- Amplitude

Brown et al '83, Grosz & Hirschberg'92, Hirschberg &
Nakatani '96

- Contour

Brown et al '83, Woodbury '87, Swerts et al '92

Finding Sentence and Topic/Story Boundaries in ASR Transcripts

- Shriberg et al 2000
- Text-based segmentation is fine...if you have reliable text
- Could prosodic cues perform as well or better at sentence and topic segmentation in ASR transcripts? – more robust? – more general?
- Goal: identify sentence and topic boundaries at ASR-defined word boundaries
 - CART decision trees and LM
 - HMM combined prosodic and LM results

Features

- Trained/tested on Switchboard and Broadcast News
- For each potential boundary location:
 - Pause at boundary (raw and normalized by speaker)
 - Pause at word before boundary (is this a new 'turn' or part of continuous speech segment?)
 - Phone and rhyme duration (normalized by inherent duration) (phrase-final lengthening?)
 - F0 (smoothed and stylized): reset, range (topline, baseline), slope and continuity
 - Voice quality (halving/doubling estimates as correlates of creak or glottalization)
 - Speaker change, time from start of turn, # turns in conversation and gender

Sentence Segmentation Results

- Prosodic only model
 - Better than LM for BN
 - Worse (on hand transcription) and same (for ASR transcript) on SB
 - Slightly improves LM on SB
- Useful features for BN
 - Pause at boundary, turn change/no turn change, f0 diff across boundary, rhyme duration
- Useful features for SB
 - Phone/rhyme duration before boundary, pause at boundary, turn/no turn, pause at preceding word boundary, time in turn

Topic Segmentation Results (BN only):

- Useful features
 - Pause at boundary, f0 range, turn/no turn, gender, time in turn
- Prosody alone better than LM
- Combined model improves significantly

Story Segmentation on BN

- Rosenberg et al '07
- Goal: Divide each show into homogenous regions, each about a single topic
 - Task: Focused Q/A
 - Issue: What unit of analysis should we use in identifying potential boundaries?

TDT-4 Corpus

- English: 312.5 hours, 250 broadcasts, 6 shows
- Arabic: 88.5 hours, 109 broadcasts, 2 shows
- Mandarin: 109 hours, 134 broadcasts, 3 shows
- Manually annotated story boundaries
- ASR Hypotheses
- Speaker Diarization Hypotheses

Approach

- Identify set of segments which define:
 - Unit of analysis
 - Candidate boundaries
- Classify each candidate boundary based on features extracted from segments
 - C4.5 Decision Tree
 - Model each show-type separately
 - E.g. CNN “Headline News” and ABC “World News Tonight” have distinct models
 - Evaluate using WindowDiff with $k=100$

Segment Boundary Modeling Features

- Acoustic
 - Pitch & Intensity
 - speaker normalized
 - min, mean, max, stdev, slope
 - Speaking Rate
 - vowels/sec, voiced frames/sec
 - Final Vowel, Rhyme Length
 - Pause Length
- Lexical
 - TextTiling scores
 - LCSeg scores
 - Story beginning and ending keywords
- Structural
 - Position in show
 - Speaker participation
 - First or last speaker turn?

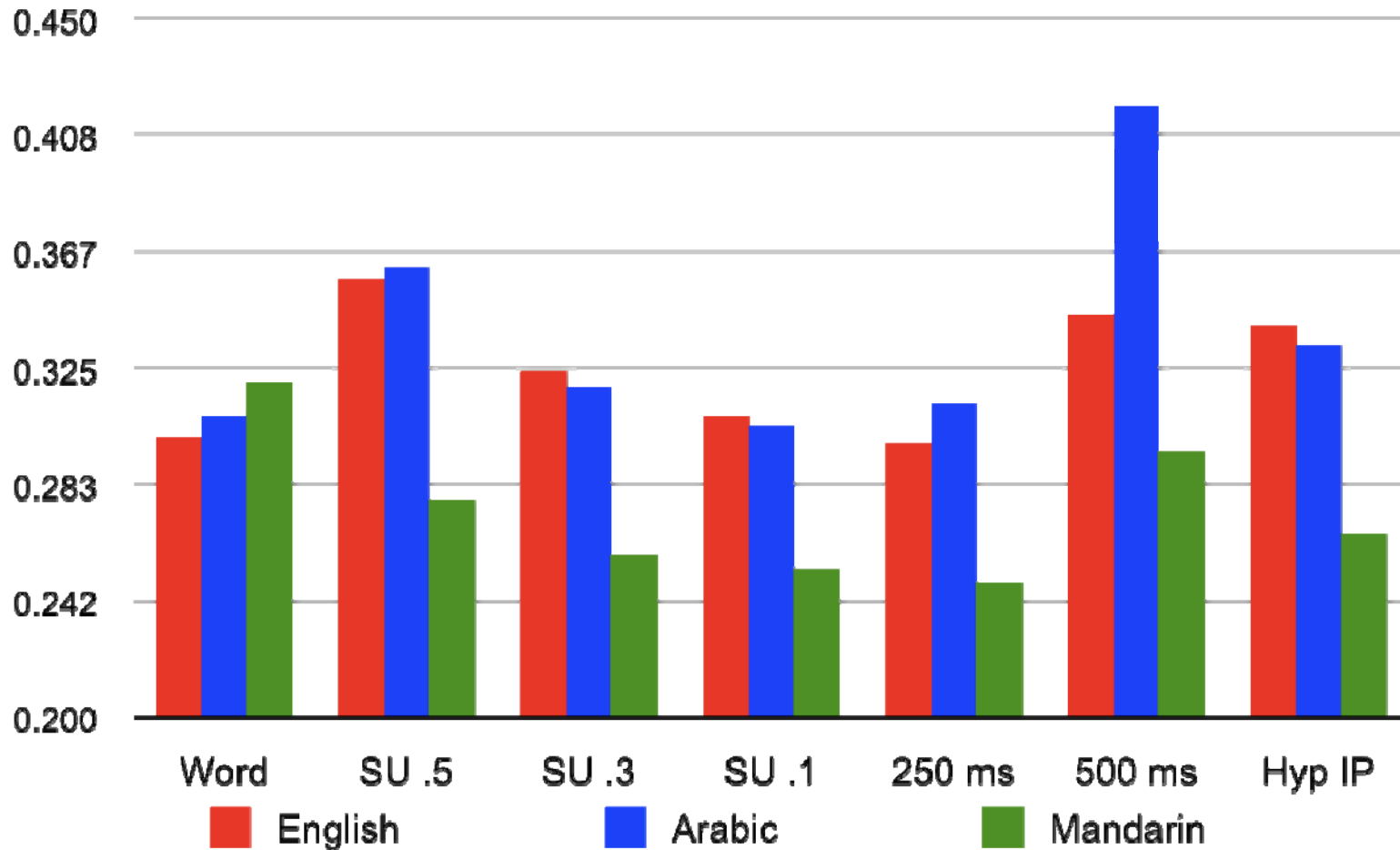
Input Segmentations

- ASR Word boundaries
 - No segmentation baseline
- Hypothesized Sentence Units
 - Boundaries with 0.5, 0.3 and 0.1 confidence thresholds
- Pause-based Segmentation
 - Boundaries at pauses over 500ms and 250ms
- Hypothesized Intonational Phrases

Hypothesizing Intonational Phrases

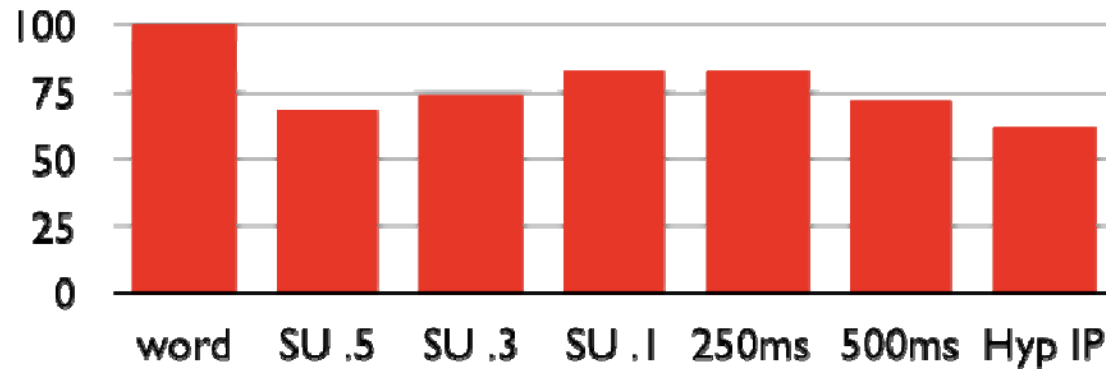
- ~30 minutes manually annotated ASR BN from reserved TDT-4 CNN show.
 - Phrase was marked if a phrase boundary occurred since the previous word boundary.
- C4.5 Decision Tree
- Pitch, Energy and Duration Features
 - Normalized by hypothesized speaker id and surrounding context
- 66.5% F-Measure (p=.683, r=.647)

Story Segmentation Results

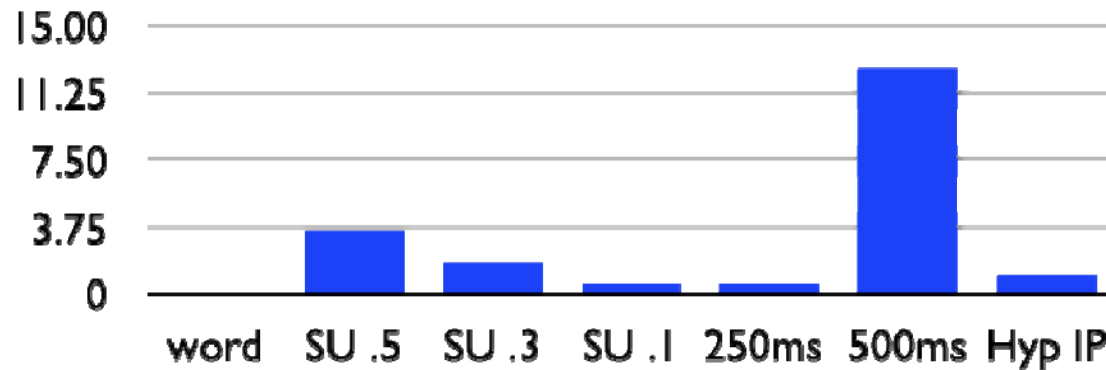


Input Segmentation Statistics

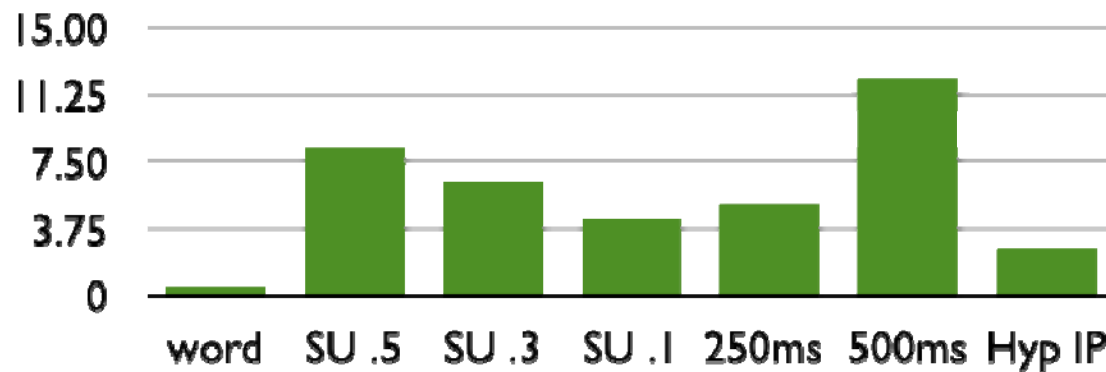
Exact Story
Boundary
Coverage (pct.)



Mean Distance to
Nearest True
Boundary (words)



Segment to True
Boundary Ratio






4/6/2011

Results

- **Best Performance:**
 - Low threshold (0.1) sentences
 - Short pause (250ms) segmentation
 - Hyp. IPs perform better than sentences.
 - Would increased SU, IP accuracy improve story segmentation?
- **External evaluation: impact on IR and MT performance.**

Disfluencies and Self-Repairs

- Spontaneous speech is ‘ungrammatical’
 - every 4.6s in radio call-in (Blackmer & Mitton ‘91)
 - hesitation: *Ch- change strategy.* 
 - filled pause: *Um Baltimore.* 
 - self-repair: *Ba- uh Chicago.* 
- A big problem for speech recognition
 - Ch- change strategy.* --> to D C D C today ten fifteen.
 - Um Baltimore.* --> From Baltimore ten.
 - Ba- uh Chicago.* --> For Boston Chicago.

Disfluencies as 'Noise'

- For people
 - Repairs as replanning events
 - Repairs as attention-getting devices (taking the turn)
- For parsers
- For speech recognizers

What's the Alternative?

- Modeling disfluencies
 - Filled pauses
 - Self-repairs
 - Hesitations
- Detecting disfluencies explicitly
 - Why is this hard?
 - Distinguishing them from 'real' words (uh vs. a)
 - Distinguishing them from 'real' noise

Self-Repairs

- Hindle '83:
 - When people produce disfluent speech and correct themselves....
 - They leave a trail behind
 - Hearers can compare the fluent finish with the disfluent start
 - This is a bad – a disastrous move
 - 'a/DET bad/ADJ'/'a/DET disastrous/ADJ'
 - To determine what to 'replace' with what
 - Corpus: interview transcripts with correct p.o.s. assigned

The '*Edit Signal*'

- How do Hearers know what to keep and what to discard?
- Hypothesis: Speakers signal an upcoming repair by some acoustic/prosodic **edit signal**
 - Tells hearers where the disfluent portion of speech ends and the correction begins
 - Reparandum – edit signal – repair
- What I {uh,I mean, I-,...} what I said is
- If there is an edit signal, what might it be?
 - Filled pauses
 - Explicit words
 - Or some 'non-lexical' acoustic phenomena

Categories of Self Repairs

- Same surface string
Well if they'd * if they'd...
- Same part-of-speech
I was just that * the kind of guy...
- Same syntactic constituent
I think that you get * it's more strict in Catholic schools
- Restarts are completely different...
I just think * Do you want something to eat?

Hindle Category Distribution for 1 Interview 1512 sentences, 544 repairs

Category	N	%
Edit Signal Only	128	24%
Exact Match	161	29%
Same POS	47	9%
Same Syntactic Constituent	148	27%
Restart	32	6%
Other	28	5%

But *is* there an Edit Signal?

- Definition: a *reliable* indicator that divides the reparandum from the repair
- In search of the edit signal: [RIM Model](#) of Self-Repairs (Nakatani & Hirschberg '94)
 - Reparandum, Disfluency Interval (Interruption Site), Repair
- ATIS corpus
 - 6414 turns with 346 (5.4%) repairs, 122 speakers, hand-labeled for repairs and prosodic features

xwaves Multidimensional Signal Display, Version 5.3.1 (xwaves)

OBJECT name: selina [PAUSE] [CONTINUE] [xwaves MANUAL] [QUIT!]
INPUT file: selina.d OUTPUT file: foo1
Overlay name: Attach: [xspectrum] [xlabel] [xchart]
COMMAND (or @file): @/tmp/transcribe3595431

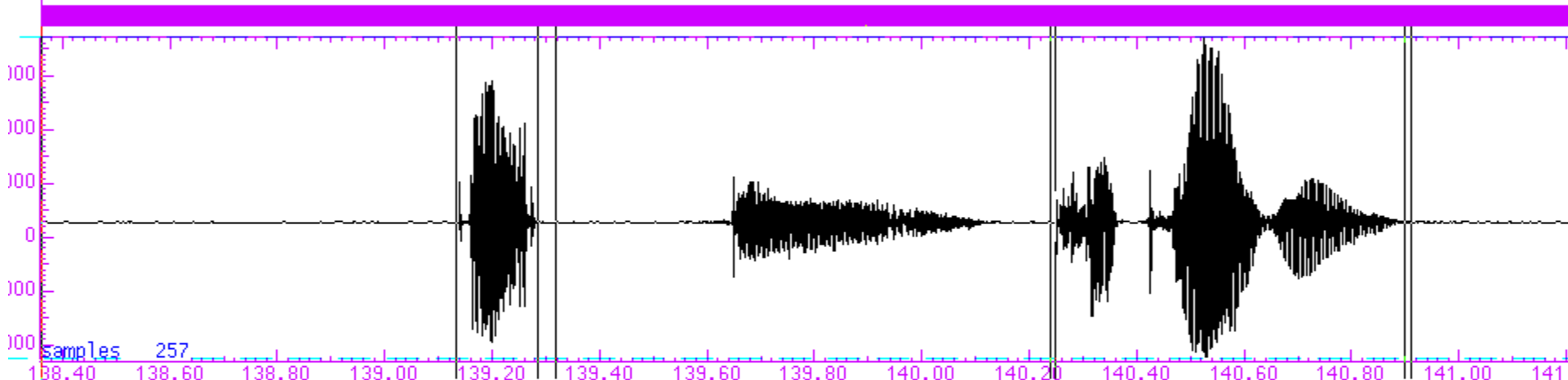
Labeler 5.3.1 (xlabel)

Label File: selina.words
Object: selina [xlabel manual]
Active fields: 1
Label Menu File: /s/menus/labelmenu

selina.d (S.F.: 8000.0) {left:up/down move mid:play between marks right:menu}

Time:141.21436sec

D: 3.22487 L:138.36100 R:141.58588 (F: 0.31)



selina

Ba uh Chicago

<reparandum IS disfluency interval> repair>
reparandum> <repair

selina.rim T:141.21436 INSERT MODE

Lexical Class of Word Fragments Ending Reparandum

Lexical Class	N	%
Content words	128	43%
Function words	14	5%
?	156	52%

Length of Fragments at End of Reparandum

Syllables	N	%
0	119	40%
1	153	51%
2	25	8%
3	1	.3%

Length in Words of Reparandum

Length	Fragment Repairs (N=280)		Non-Fragment Repairs (N=102)	
1	183	65%	53	52%
2	64	23%	33	32%
3	18	6%	9	9%
4	6	2%	2	2%
5 or more	9	3%	5	5%

Type of Initial Phoneme in Fragment

Class of First Phoneme	% of All Words	% of All Fragments	% of 1-Syl Fragments	% of 1-C Fragments
Stop	23%	23%	29%	12%
Vowel	25%	13%	20%	0%
Fricative	33%	44%	27%	72%
Nasal/glid e/liquid	18%	17%	20%	15%
H	1%	2%	4%	1%
Total N	64,896	298	153	119

Presence of Filled Pauses/Cue Phrases

	FP/Cue Phrases	Unfilled Pauses
Fragment	16	264
Non-Fragment	20	82

Duration of Pause

	Mean	SDev	N
Fluent Pause	513ms	676ms	1186
DI	334ms	421ms	346
Fragment	289ms	377ms	264
Non-Fragment	481ms	517ms	82

Is There an Edit Signal?

- Findings:
 - Reparanda: 73% end in fragments, 30% in glottalization, co-articulatory gestures
 - DI: pausal duration differs significantly from fluent boundaries, small increase in f0 and amplitude
- Speculation: articulatory disruption
- Are there edit *signals*?

With or Without an Edit Signal, How Might Hearers/Machines Process Disfluent Speech?

- Parsing-based approaches: (Weischedel & Black '80; Carbonell & Hayes '83; Hindle '83; Fink & Biermann '86):
 - If 2 constituents of identical semantic/syntactic type are found where grammar allows only one, delete the first
 - Use an 'edit signal' or explicit words as cues
 - Select the minimal constituent
- Pick up the blue- green ball.

- Results: Detection and correction
 - Trivial (edit signal only): 128 (24%)
 - Non-trivial: 388 (71%)

Pattern-matching approaches (Bear et al '92)

- Find candidate self-repairs using lexical matching rules
 - Exact repetitions within a window
I'd like a a tall latte.
 - A pair of specified adjacent items
The a great place to visit.
 - 'Correction phrases'
That's the well uh the Raritan Line.
- Filter using syntactic/semantic information
That's what I mean when I say it's too bad.

Distribution of Reparanda

- 10,718 utterances
- Of 646 repairs:
 - Most nontrivial repairs (339/436) involve matched strings of identical words
 - Longer matched string
 - More likely a repair
 - More words between matches
 - Less likely repair
- Distribution of reparanda by ***length in words*** ----->

Len	N	%
1	376	59%
2	154	24%
3	52	8%
4	25	4%
5	23	4%
6+	16	3%

- Detection results:
 - 201 ‘trivial’ (fragments or filled pauses)
 - Of 406 remaining:
 - Found 309 correctly (76% Recall)
 - Hypothesized 191 incorrectly (61% Precision)
 - Adding ‘trivial’: 84% Recall, 82% Precision
- Correcting is harder:
 - Corrects all ‘trivial’ but only 57% of correctly identified non-trivial

Machine Learning Approaches (Nakatani & Hirschberg '94)

- CART prediction: 86% precision, 91% recall
 - Features: Duration of interval, presence of fragment, pause filler, p.o.s., lexical matching across DI
 - Produce rules to use on unseen data
 - But...requires hand-labeled data

State of the Art (Liu et al 2002,2005)

- Detecting the Interruption Point (IP) using acoustic/prosodic and lexical features
- Features:
 - Normalized duration and pitch features
 - Voice quality features:
 - Jitter: perturbation in the pitch period
 - Spectral Tilt: overall slope of the spectrum
 - Open Quotient: ratio of time vocal folds open/total length of glottal cycle
 - Language Models: words, POS, repetition patterns

	I	hope	to	have	to	have	
	NP	VB	PREP	VB	PREP	VB	
X	X	Start	Orig2	IP	Rep	End	

- Corpus:
 - 1593 Switchboard conversations, hand-labeled
 - Downsample to 50:50 IP/not since otherwise baseline is 96.2% (predict no IP)
- Results:
 - Prosody alone produces best results on downsampled data (Prec. 77%, Recall 76%)

- IP Detection: Precision/Recall
 - Prosody+Word LM+POS LM does best on non-downsampled (Prec.57%, Recall 81%)
- IP Detection: Overall accuracy
 - Prosody alone on reference transcripts (77%) vs. ASR transcripts (73%) -- ds
 - Word LM alone on reference transcripts (98%) vs ASR transcripts (97%) – non-ds
- Finding reparandum start:
 - Rule-based system (Prec. 69%, Recall 61%)
 - LM (Prec. 76%, Recall 46%)
- Have we made progress?

IP Detection Results

- Downsampled
 - Chance - - 50 (Acc)
 - Prosody 75.81 77.26 76.75 (P,R,A)
- Non-downsampled
 - Chance 0 - 96.62 (A)
 - Prosody 0 - 96.62 (A)
 - Word-LM 55.47 79.33 98.01 (P,R,A)
 - POS-LM 36.73 65.75 97.22
 - Word-LM+Prosody 58.27 78.37 98.05
 - **Word-LM+ Prosody+ POS-LM 56.76 81.25 98.10**

Next Class

- Spoken Dialogue Systems