

CS 224S LING 281 Speech Recognition and Synthesis

Lecture 16: Metadata: disfluencies and boundaries
Dan Jurafsky

CS 224S W2006 1

Outline

- Disfluencies
- Characteristics of disfluencies
- Detecting disfluencies
- MDE bakeoff
- Fragments

CS 224S W2006 2

Disfluencies

the . [exhale] . . . [inhale] . . . [uh] does American airlines . offer any . one way flights . [uh] one way fares, for one hundred and sixty one dollars
[mm] i'd like to leave i guess between [um] . [smack] . five o'clock no, five o'clock and [uh], seven o'clock . P M
around, four, P M
all right, [throat.clear] . . . i'd like to know the . give me the flight . times . in the morning . for September twentieth . nineteen ninety one
[uh] one way
[uh] seven fifteen, please
on United airlines . . . give me, the . . time . . from New York . [smack] . to Boise-, to . I'm sorry . on United airlines . [uh] give me the flight, numbers, the flight times from . [uh] Boston . to Dallas

Figure 9.5 Some sample spoken utterances from users interacting with the ATIS system.

CS 224S W2006 3

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

CS 224S W2006 4

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
- Might help in language modeling
 - Disfluencies might occur at particular positions (boundaries of clauses, phrases)
- Annotating them helps readability
- Disfluencies cause errors in nearby words

CS 224S W2006 5

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)

CS 224S W2006 6

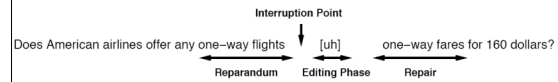
Prosodic characteristics of disfluencies

- 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

DS 2245 W2006 7

Syntactic Characteristics of Disfluencies

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



DS 2245 W2006 8

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

DS 2245 W2006 9

Um versus uh: delays (Clark and Fox Tree)

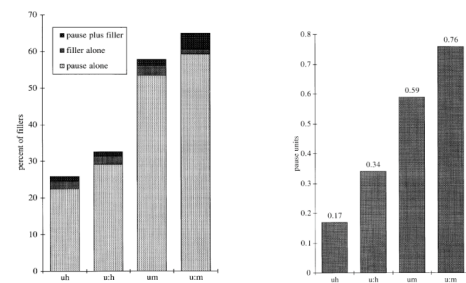


Fig. 1. Percent of fillers followed by delays (LL corpus).

Fig. 2. Mean length of pauses after fillers (LL corpus).

Utterance Planning

- 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?

DS 2245 W2006 11

Delays at different points in phrase

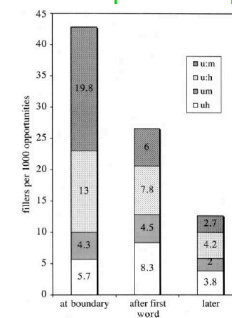


Fig. 5. Rates of uh and um at three position in tone units (LL corpus).

Disfluencies in language modeling

- Should we “clean up” disfluencies before training LM (i.e. skip over disfluencies?)
 - Filled pauses
 - Does United offer any [uh] one-way fares?
 - Repetitions
 - What what are the fares?
 - Deletions
 - Fly to Boston from Boston
 - Fragments (we’ll come back to these)
 - I want fl- flights to Boston.

CS 224S W2006 14

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

CS 224S W2006 14

Change in Perplexity when Filled Pauses (FP) are removed

- LM Perplexity goes up at following word:

Position	UH	UH+1	UH+2	UM	UM+1	UM+2	non-FP	overall
Baseline	39.0	223.5	89.5	174.9	36.7	71.9	103.4	101.9
FP model	39.0	281.5	91.4	175.5	73.4	68.2	103.4	103.3
#events	502	502	373	188	188	84		19426

- Removing filled pauses makes LM worse!!
- I.e., filled pauses seem to help to predict next word.
- Why would that be?

Stolcke and Shriberg 1996

CS 224S W2006 15

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 [uh] there’s not time to do it all
- “there’s” is very likely to start a sentence
- So $P(\text{there’s|uh})$ is better estimate than $P(\text{there’s|doing})$

CS 224S W2006 16

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”

Position	UH+1	UM+1
Baseline	849.0	437.4
FP model	606.2	361.7

CS 224S W2006 17

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!

CS 224S W2006 18

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

CS 224S W2006

19

WER reductions from modeling disfluencies+background events

- Rose and Riccardi (1999)

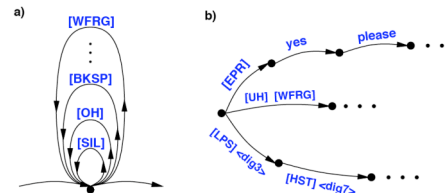


Figure 1: a) Inclusion of all labeled background events (LBEs) in a single "between-word" loop. b) Portion of phrase-based LM trained from LBE annotated text.

CS 224S W2006

20

HMIHY Background events

- Out of 16,159 utterances:

Filled Pauses	7189
Word Fragments	1265
Hesitations	792
Laughter	163
Lipsmack	2171
Breath	8048
Non-Speech Noise	8834
Background Speech	3585
Operator Utt.	5112
Echoed Prompt	5353

CS 224S W2006

21

Phrase-based LM

- "I would like to make a collect call"
- "a [wfrag]"
- <dig3> [brth] <dig3>
- "[brth] and I"

CS 224S W2006

22

Rose and Riccardi (1999) Modeling "LBEs" does help in WER

ASR Word Accuracy			
System Configuration		Test Corpora	
HMM	LM	Greeting (HMIHY)	Card Number
Baseline	Baseline	58.7	87.7
LBE	Baseline	60.8	88.1
LBE	LBE	60.8	89.8

CS 224S W2006

23

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

CS 224S W2006

24

Detection of disfluencies

- Nakatani and Hirschberg
- Decision tree at wi-wj boundary
 - pause duration
 - Word fragments
 - Filled pause
 - Energy peak within wi
 - Amplitude difference between wi and wj
 - F0 of wi
 - F0 differences
 - Whether wi accented
- Results:
 - 78% recall/89.2% precision

CS 224S W2006 25

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%

CS 224S W2006 26

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

CS 224S W2006 27

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

CS 224S W2006 28

NL Constraints

- Syntax Only
 - Precision: 96%
- Syntax and Semantics
 - Correction: 70%

CS 224S W2006 29

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 2005 CS 224S W2006 30

Kinds of disfluencies

- Repetitions
 - I * I like it
- Revisions
 - We * I like it
- Restarts (false starts)
 - It's also * I like it

CS 224S W2006 11

MDE transcription

- 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 ./

CS 224S W2006 12

MDE Labeled Corpora

	CTS	BN
Training set (words)	484K	182K
Test set (words)	35K	45K
STT WER (%)	14.9	11.7
SU %	13.6	8.1
Edit word %	7.4	1.8
Filler word %	6.8	1.8

CS 224S W2006 13

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

CS 224S W2006 14

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

CS 224S W2006 15

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

CS 224S W2006 16

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

CS 224S W2006 37

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

CS 224S W2006 38

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

CS 224S W2006 39

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

CS 224S W2006 40

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
- 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)

CS 224S W2006 41

Why fragments are important

- Frequent enough to be a problem:
 - Only 1% of words/3% of sentences
 - But if miss fragment, likely to get surrounding words wrong (word segmentation error).
 - So could be 3 or more times worse (3% words, 9% of sentences): problem!
- Useful for finding other repairs
 - In 40% of SRI-ATIS sentences containing fragments, fragment occurred at right edge of long repair
 - 74% of ATT-ATIS reparanda ended in fragments
- Sometimes are the only cue to repair
 - "leaving at <seven> <fif-> eight thirty"

CS 224S W2006 42

How fragments are dealt with in current ASR systems

- In training, throw out any sentences with fragments
- In test, get them wrong
- Probably get neighboring words wrong too!
- !!!!!

CS 224S W2006 44

Cues for fragment detection

- 49/50 cases examined ended in silence >60msec; average 282ms (Bear et al)
- 24 of 25 vowel-final fragments glottalized (Bear et al)
 - Glottalization: increased time between glottal pulses
- 75% don't even finish the vowel in first syllable (i.e., speaker stopped after first consonant) (O'Shaughnessy)

CS 224S W2006 44

Cues for fragment detection

- Nakatani and Hirschberg (1994)
- Word fragments tend to be content words:

Lexical Class	Token	Percent
Content	121	42%
Function	12	4%
Untranscribed	155	54%

CS 224S W2006 45

Cues for fragment detection

- Nakatani and Hirschberg (1994)
- 91% are one syllable or less

Syllables	Tokens	Percent
0	113	39%
1	149	52%
2	25	9%
3	1	0.3%

CS 224S W2006 46

Cues for fragment detection

- Nakatani and Hirschberg (1994)
- Fricative-initial common; not vowel-initial

Class	% words	% frags	% 1-C frags
Stop	23%	23%	11%
Vowel	25%	13%	0%
Fric	33%	45%	73%

CS 224S W2006 47

Liu (2003): Acoustic-Prosodic detection of fragments

- Prosodic features
 - Duration (from alignments)
 - Of word, pause, last-rhyme-in word
 - Normalized in various ways
 - F0 (from pitch tracker)
 - Modified to compute stylized speaker-specific contours
 - Energy
 - Frame-level, modified in various ways

CS 224S W2006 48

Liu (2003)

- Use Switchboard 80%/20%
- Downsampled to 50% frags, 50% words
- Generated forced alignments with gold transcripts
- Extract prosodic and voice quality features
- Train decision tree

CS 224S W2006 49

Liu (2003) results

- Precision 74.3%, Recall 70.1%

		hypothesis	
		complete	fragment
reference	complete	109	35
	fragment	43	101

CS 224S W2006 50

Liu (2003) features

- Features most queried by DT

Feature	%
jitter	.272
Energy slope difference between current and following word	.241
Ratio between F0 before and after boundary	.238
Average OQ	.147
Position of current turn	0.084
Pause duration	0.018

CS 224S W2006 51

Liu (2003) conclusion

- Very preliminary work
- Fragment detection is good problem that is understudied!

CS 224S W2006 52

Liu (2003): Acoustic-Prosodic detection of fragments

- Voice Quality Features
 - Jitter
 - A measure of perturbation in pitch period
 - Praat computes this
 - Spectral tilt
 - Overall slope of spectrum
 - Speakers modify this when they stress a word
 - Open Quotient
 - Ratio of times in which vocal folds are open to total length of glottal cycle
 - Can be estimated from first and second harmonics
 - Creaky voice (laryngealization) vocal folds held together, so short open quotient

CS 224S W2006 53

Summary

- Disfluencies
- Characteristics of disfluencies
- Detecting disfluencies
- MDE bakeoff
- Fragments

CS 224S W2006 54