Processing Spoken Language

Mari Ostendorf
University of Washington

Research sponsored at various times by Google, Amazon, Tencent, NSF & DARPA

Growth in Speech Interfaces & Online Content

- People like spoken communication
  - Video & podcast content online is rapidly growing
  - It is often easier to talk than type (driving, cooking, …)
  - Speech technology makes information accessible
- Speech recognition & synthesis have advanced!

What's next? Spoken language processing!

- Speech & NLP have both advanced with many of the same neural models
- Combining speech + language enables
  - Speech interfaces for more complex tasks
  - Making speech & video as accessible as text
- Many NLP tasks apply to spoken language
  - Summarization, translation
  - Information extraction, question answering
  - Chat and assistant dialogs

Talk Overview

- Spoken language ≠ noisy verbalized text
- General challenges in applying NLP to spoken language
- Challenges of conversational systems
- Ethical & social issues

Spoken Language ≠ Noisy Verbalized Text

- Imperfect speech recognition
- Disfluencies (fillers, repetitions, self-corrections)
- Style differences
- Extra information communicated in emphasis & intonation

ASR is imperfect (even now)

yeah can i get my butt hampshire suspense are there was a cough sure stop

No problem. Let me think. How about we chat about…

…

cause does that you're gonna state that's cool

I'm happy you liked that.

- Missed/false detection of speech
- Word recognition errors
- No sentence segmentation

Disfluencies

How most people talk in conversations

~~What did~~ … what'd I do ~~the o~~ ... the other day? ~~It was~~ ... oh ~~the the then th th~~ the pork ribs and ~~the~~ … bunch of Korean food and stuff … ~~but~~ … yeah … I don't know.

What we take away

What did I do the other day? Oh, the pork ribs and bunch of Korean food and stuff. Yeah, I don't know.

… as do justices and lawyers

Yes. ~~***That***~~ that is ~~***there there***~~ ***uh uh*** there are two arguments about the risk of corruption. At the moment the argument that I'm talking about is ~~***that the party is a means that that to that***~~ that the ***um*** contribution limits on individual donors are justified as a means of preventing ***uh*** corruption …

~~***It would have to be***~~ I would think a reasonable standard ~~***is***~~ would have to be …

… and children

~~***because it's***~~ because it's 'citing ~~***it's i-and***~~ and also she must do it because if she want to do the same thing as me then if she want to do it then she have to … ***um uh*** you just have to be nice and act like this excuse me ***um mm*** could I paint with you this evening … if she doesn't ~~***then then then s- hen y- then then you c-***~~ you can't yell

(from Alwan & Bailey, UCLA)

Style Differences

Portuguese cuisine was first recorded in the seventeenth century, with regional recipes establishing themselves in the nineteenth century. Culinária Portuguesa, by António-Maria De Oliveira Bello, better known as Olleboma; was the first 'Portuguese-only' recipe book published in 1936. Despite being relatively restricted to an Atlantic, Celtic sustenance, the Portuguese cuisine also has …

Wikipedia

Blog Post

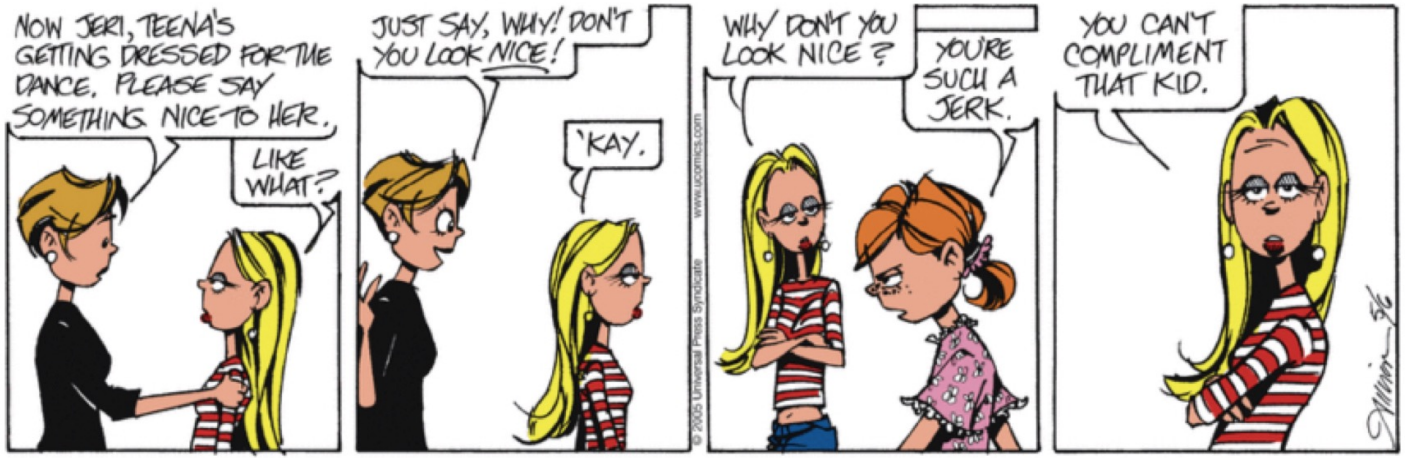
I am so excited to share a little bit of my world with the world!
Join me throughout your week to peek at my creations… look in on some Portuguese-American culture… pick up a new recipe for a weeknight dinner or plan a fun party menu… find a new fun way to have fun with your kids… or just look at some pretty pictures and maybe get inspired to create your own fun work of art!

A: yeah and we've been also doing more cooking which is fun
B: yeah that's that's true
A: like the wha- how do you say it again? uh bac- bacanau?
B: oh bacalhao
A: bac- bacalhao
B: yeah that was that was a recipe that I wanted to do for quite some time

Conversation

Prosody

It’s not what you say, but how you say it.



PreTeena by Allison Barrows -- May 6, 2005

Found in the Language Log archives.

Prosody for human-centered interactions

- Is the user talking to the device? …finished with the turn?
- Sentence segmentation, statement vs. question
- Multiple meanings of “yeah”
  - Positive answer
  - Backchannel, turn taking
  - Sarcastic (negative)
- Meaning of interjections: “oh”

Did you know that …?

yeah I did not

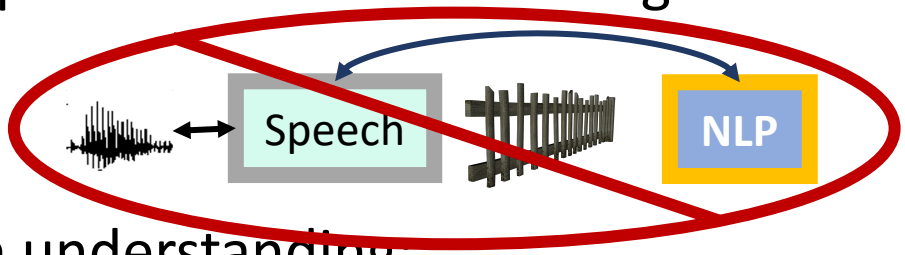
I get it I remember Of course Unenthusiastic

Computational Modeling Challenges for Prosody

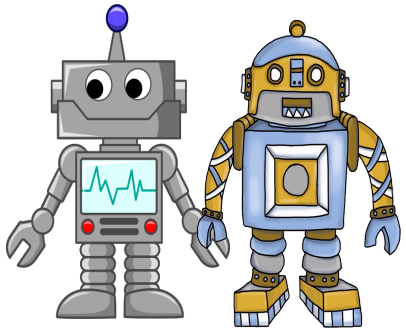
- Prosodic cues are associated with multiple phenomena at different time scales – hard to disentangle
- Differences between read and spontaneous speech
  - Prosody carries more information in spontaneous speech 
  - Much prior work on read speech 
- In speech understanding…
  - The words matter more; effort on ASR has given greater benefits
  - Prosody used in early dialog systems reflects user uncertainty
- With the major ASR advances, it's worth revisiting prosody

Marriage of Speech & Language

- To build a human-computer dialog system or process archives of conversational speech, we need speech + NLP
- SLP requires more than throwing words over a fence



- Speech understanding:
  - NLU needs ASR uncertainty and prosodic information
- Speech generation
  - NLG need to provide intent information for controlling prosody



Two Challenges in Applying NLP to Speech

- Domain mismatch is a bigger issue for speech than text
  - Spoken language data is more costly to collect than text (word transcription + meaning annotation)
  - More variation in spoken language (wording and prosody differences)
- Acoustically marked structure: Disfluencies & segmentation
  - Key challenge is integrating word and prosodic cues

NLP benefits from pre-trained word vectors

- Represent a word sequence with contextualized word embeddings
  - Learn a neural language model, hidden state represents a word in context

    RNN + word prediction objective → ELMo
    Transformer + masking objective → BERT, XLNet, RoBERTa, ...

- Use vast amounts of text to learn general-purpose embeddings, which are plugged into task-specific networks

- Domain mismatch --> fine-tune the language model (or the last layers)

Informality → Domain Mismatch Problem

- Is written text useful for parsing speech?
  - Treebanked WSJ text: limited
  - Large amount of general text: helpful with fine-tuning

Parser Training	Embeddings	F1
Text (WSJ)	General Text	77.4
Conversations (Swbd)	Conversations	91.0
Conversations	General Text	93.2
Conversations + WSJ	General Text	93.4

(Trang et al., Interspeech 2019)

- Could we do better with more unannotated speech transcripts?
  - Researchers have trained sciBERT, bioBERT, clinicalBERT, …
  - Unfortunately, there isn't (yet) a lot of accurately transcribed, publicly-available conversational speech for training a convBERT
  - Open question whether transcripts should be accurate or ASR

Acoustically Marked Structure

- Treat structure recognition as a BIO+ word tagging problem

~~*It would have to be*~~ I would think a reasonable standard ~~*is*~~ would have to be that …

BD ID ID ID ED BC IC IC IC IC IC BED C C C C O

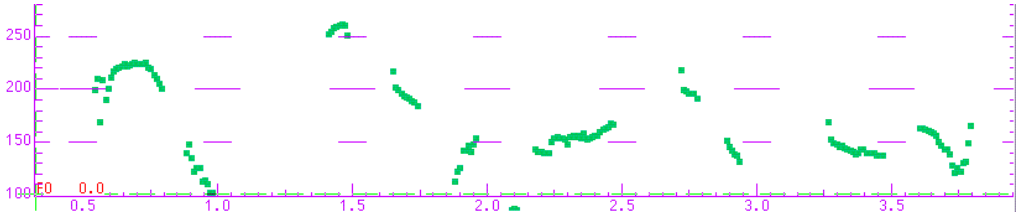
how about robots what's up with robots

BQ IQ EQ BQ IQ IQ EQ

- Detecting structure benefits from both word cues & prosody
  - Word-based prediction works well for simple cases, but less so for complex disfluencies, sentence fragments, etc.
  - Pauses are ambiguous; they mark hesitations, disfluency interruption points, and sentence/discourse structure → need prosody vectors

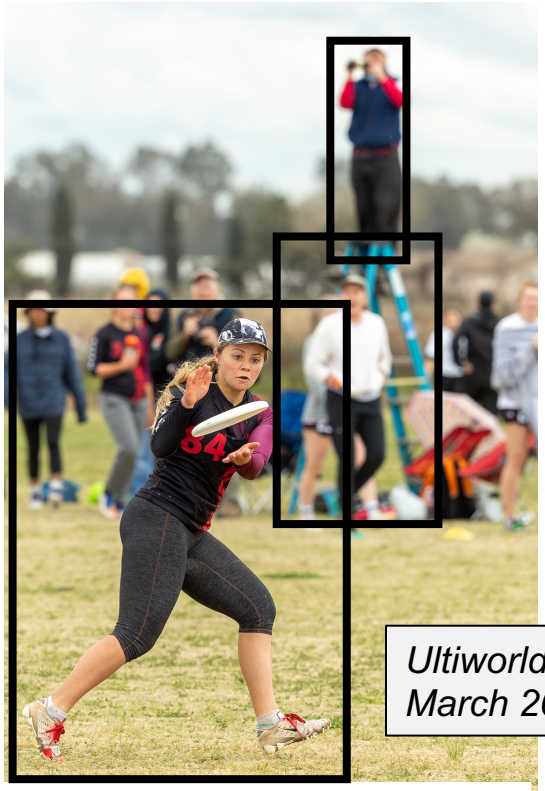
Signal Processing for Prosody

- Acoustic cues associated with prosody include fundamental frequency (F0), vocal effort, and timing (words and pauses)



- Different phenomena occur at different time scales, e.g.
  - Word emphasis: word
  - Topic change, segmentation, disfluencies: word endings and onsets
  - Asides: phrase-level differences
  - Emotion, attitude: whole segments and/or word-specific tones

Photo: Rodney Chen — UltiPhotos.com



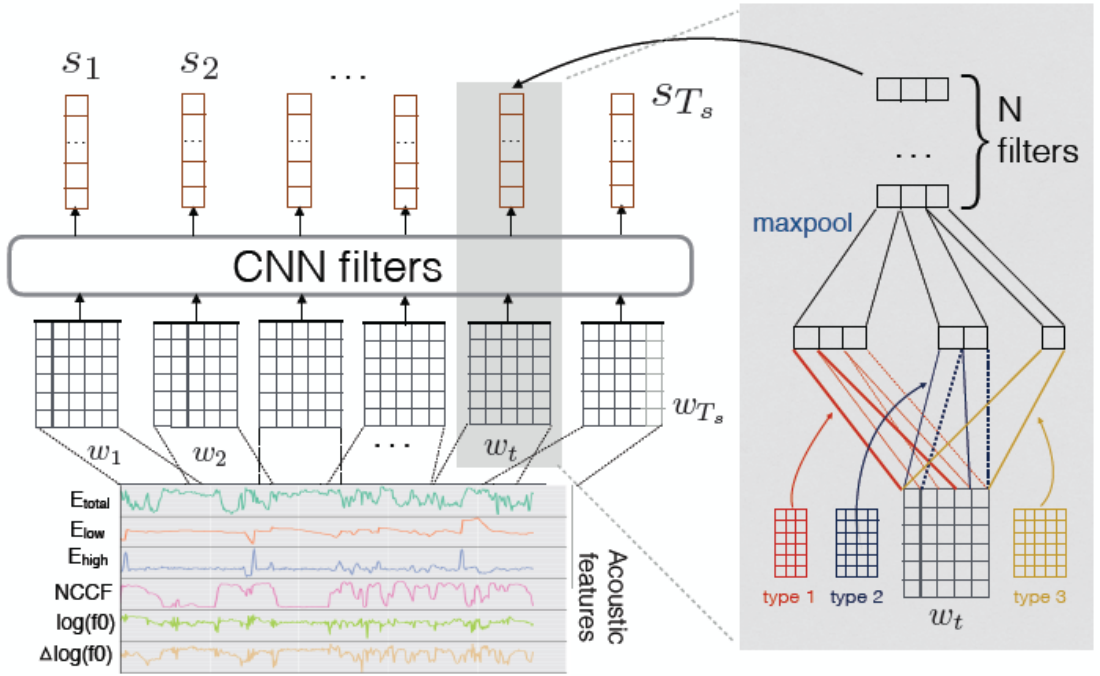
Hallie Dunham makes a catch
as Hallie's dad films in the background
atop his famous ladder.

Prosody + Text: Multimodal Analogies

- Caption & image have both complementary and redundant info (as for words & prosody)
- Caption depends on the bounding boxes; sentence meaning depends on emphasis & break location
- Standard multimodal integration issues
  - How tightly coupled modalities are
  - Modality representation learning strategy
- For sentence understanding tasks, we use
  - Word-aligned prosody vectors
  - Parallel and conditional encodings of prosody

Leveraging Prosody Features in a Neural Parser

- CNN learns F0 & energy feature functions; separate mapping for pause & duration embeddings
- Concatenate prosody and word embeddings
- End-to-end parser training
- Given sentence boundaries, parsing gains are for disfluencies & VP attachments



(Tran et al., NAACL 2018)

Contextually-Normalized Prosody Features

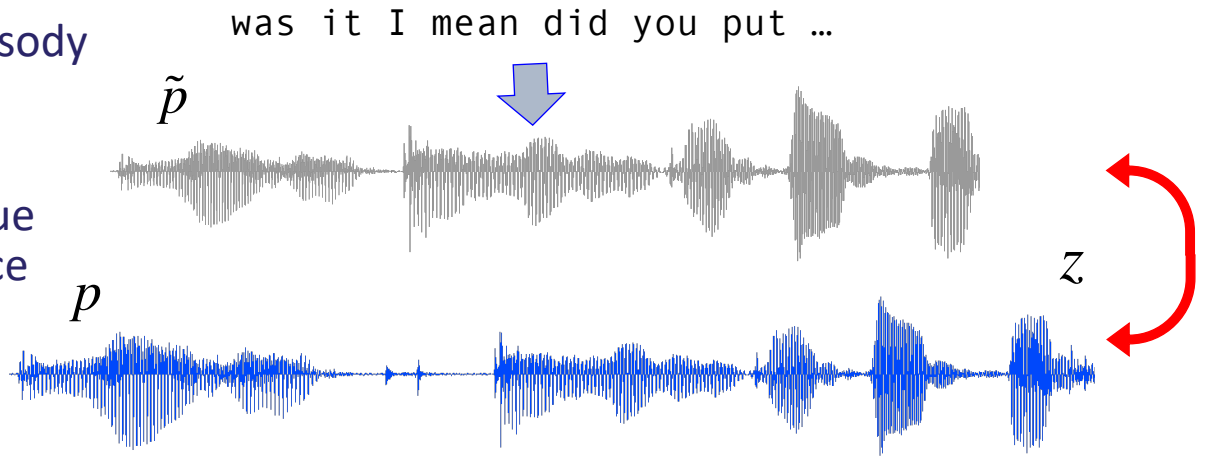
1. Given a text, predict its prosody

$$\tilde{p}_i^k \mid \bar{h}_i \sim \mathrm{N}(\mu_{i,k}, \sigma_{i,k}^2)$$

2. Compare predicted with true signal: what is the difference relative to the expected variability?

$$z_i^k = \frac{p_i^k - \mu_{i,k}}{\sigma_{i,k}}$$

3. Use as the prosody features (called *innovations*)

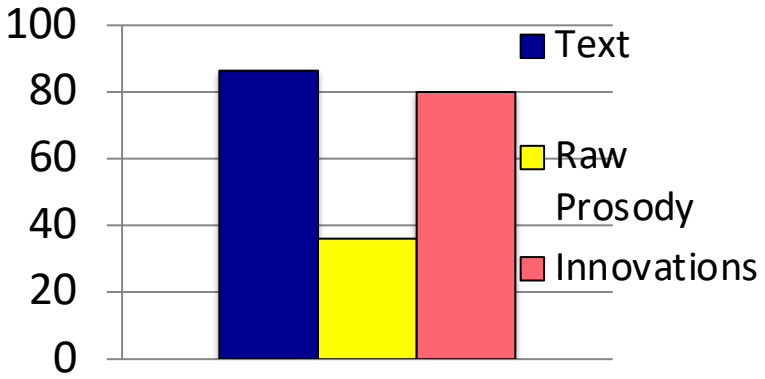


Innovations = variation that is not accounted for by the word sequence (i.e. default reading)

(Zayats et al., NAACL 2019)

Experiments in Disfluency Detection

- Task: Disfluency detection in Switchboard
- Findings:
  - Disfluency interruption points: words are longer and lower energy
  - Innovations are almost as good as text alone
  - Innovations (but raw prosody) are useful when combined with text



Examples where prosody helps:

but it‘s just you know leak leak leak everywhere

I mean [it was + it]

Conversational Speech

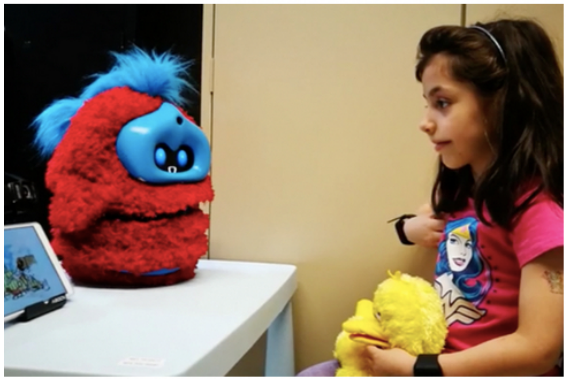
- Why conversations?
- A conversation is not a document
- Individual differences

Why conversations?

- People like spoken interactions
  - Web-based chat & FAQs have not replaced call centers
  - Informational audio is often interactive: interviews, debates, hearings, lecture Q&A
  - In the COVID-19 era, person-to-person meeting (and dinner parties) continue… virtually
- Interaction is useful for learning, problem solving, and team building

Conversations are the next frontier!

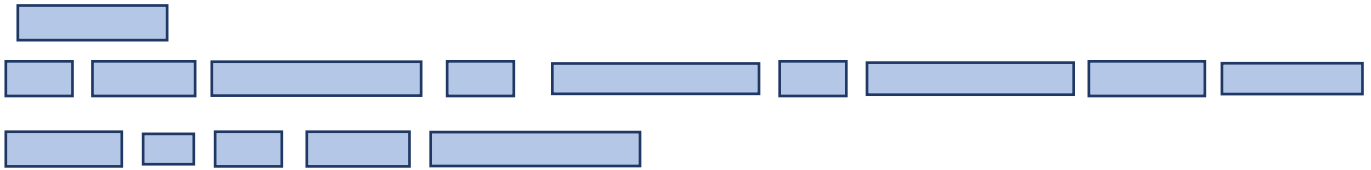
- Virtual assistants are already becoming more conversational
  (for constrained tasks)
- And there's potential for impact in many application areas
  - Call center support
  - Interactive tutorials and dialogic reading
  - Virtual companions & personal robots
  - Conversational speech translation
  - Meeting summarization
  - Medical diagnostics



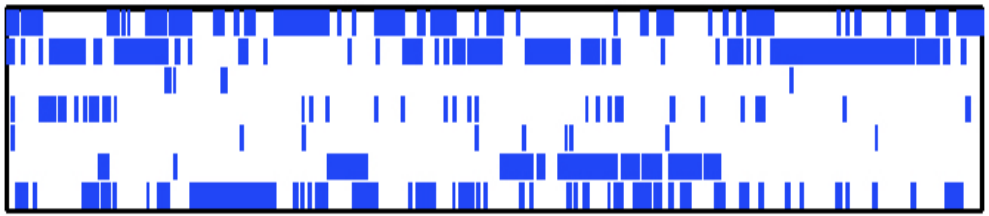
Breazeal et al. NRI project

A conversation is not a document

- A written document is a linear sequence of sentences



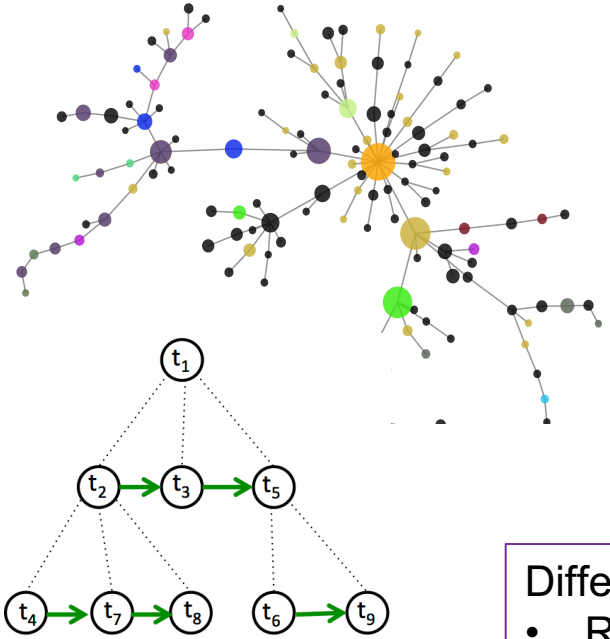
- A conversation involves multiple people that may interrupt each other, overlap, respond to earlier talkers, finish someone else's sentence, etc.



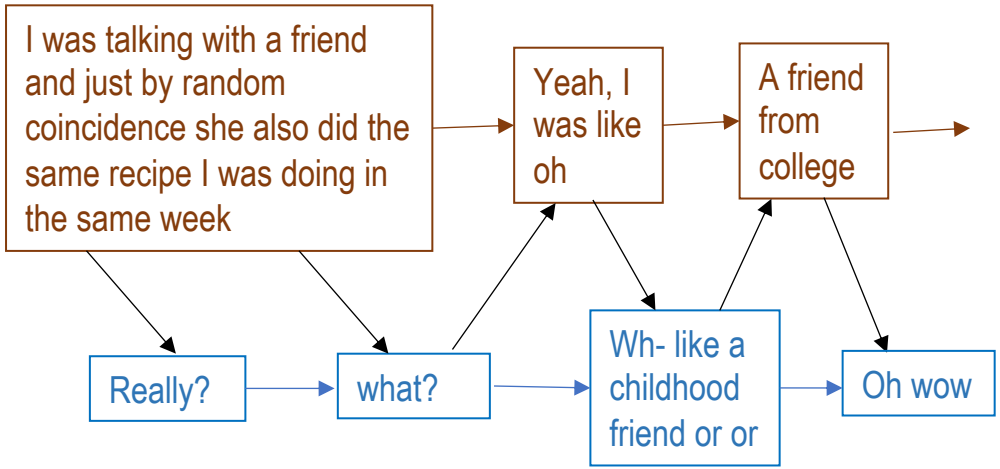
ICSI meeting

Conversations have graph structure

Online discussion

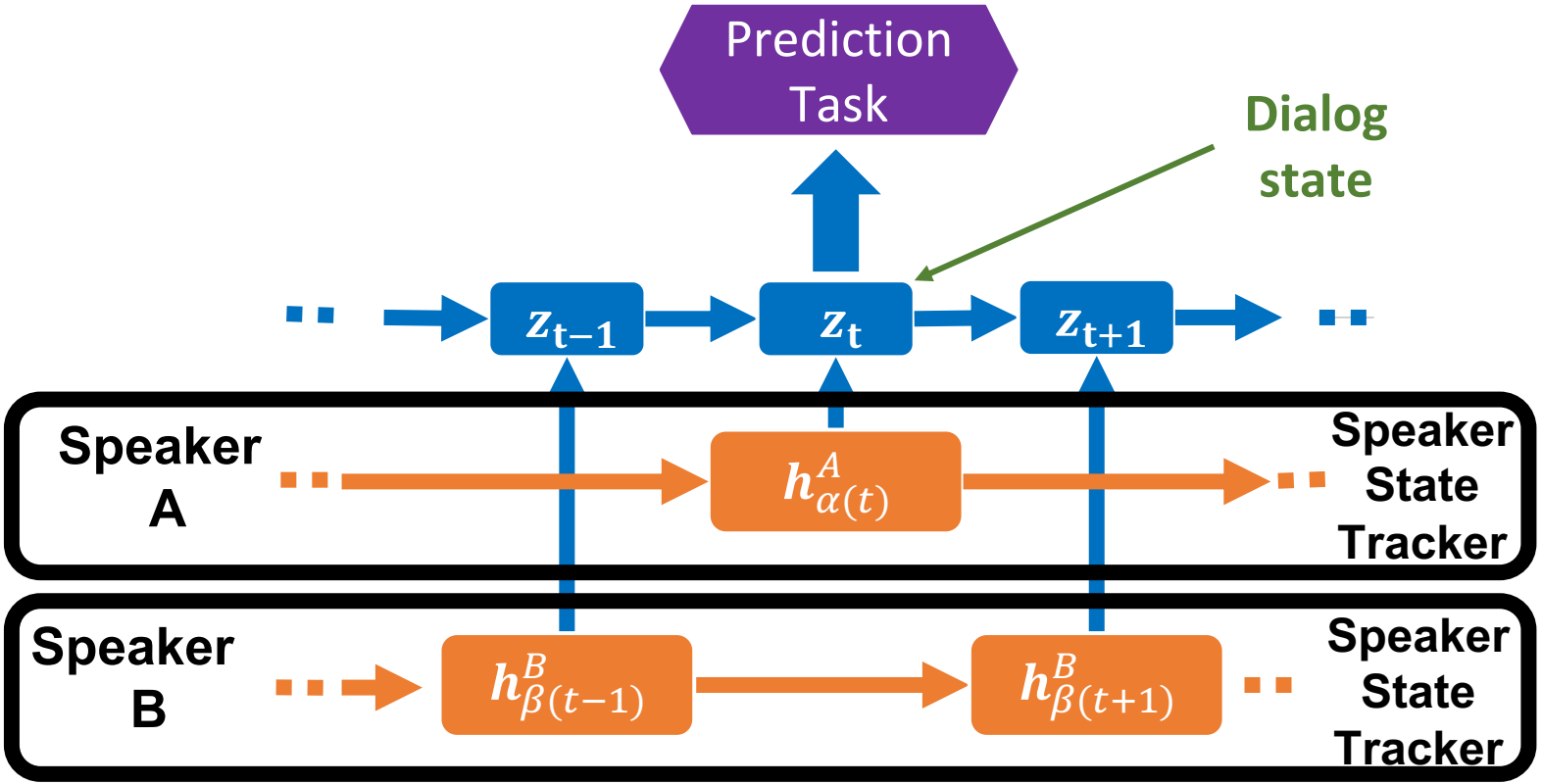


In-person conversation



Different links:
- Reply to
- Temporal

Prediction Paradigm for Dialog Tasks



Speaker Roles & Individual Differences

- People may have specific roles (host vs. guest, student vs. expert, agent vs. customer) or interaction styles

i don't know that's an interesting question and is it really true that garlic keeps vampires the wedding and what i
what are they have their long fingernails for
i think that that's probably true but i think it vampires are evil and they don't care about sustaining things for human be-...

no
yes
cool
yeah that's cool
no I didn't
No
yes

- People vary in terms of their interests, abilities, sense of humor…

Did you know that Malaysian vampires are tiny monsters that burrow into people's heads and force them to talk about cats?

Oh my god that's funny.

That's creepy.

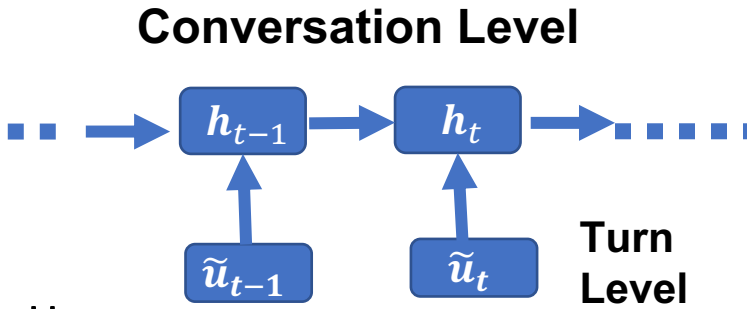
Wow that's interesting.

What the heck?

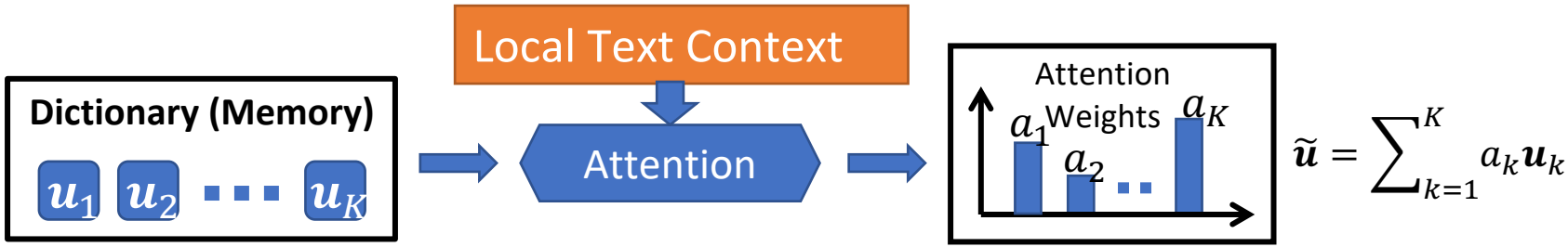
Cats are my favorite animals.

Speaker State Tracker for Dialogs

- Two types of information
  - **Short-term** modes reflected in the current utterance
  - **Long-term/carry-over** factors reflected in dialog segments



- Leverage **latent user modes** to represent an utterance



- Use next utterance (response) prediction to provide an unsupervised pre-training signal

Speaker State Tracking Studies

- Unsupervised pre-training of a speaker state sequence model benefits several tasks involving conversations
  - Reddit discussions: comment popularity prediction
  - Open-domain socialbot chat: topic acceptance prediction
  - Human-human conversation: dialog act tagging
  - Task-oriented dialog: Multi-domain state tracking
- What we learn in the speaker modes
  - Topics or topic interests
  - Sentiment, politeness
  - Course-grained dialog act

(Cheng et al., EMNLP 2017)

(Cheng et al., NAACL 2019)

(Cheng Ph.D., 2019)

Evaluation is an Open Problem

- Interactivity & individual differences pose challenges for evaluation
- For most tasks, there is no gold standard conversation
  - Once a human is in the loop, the conversation can go anywhere
  - Many responses may be acceptable
  - Single reference turn-level scoring for generation (e.g. BLEU) often don’t correlate well with human scores
- Responses that are acceptable to some are not to others

Spoken Language is Personal

- Advances in speaker recognition technology make people identifiable from their speech
- False trigger of hotword detection can expose personal info
- Advances in speech synthesis technology enable voice theft
- All these valid privacy concerns make research more difficult
  - Commercial ASR systems don't expose the audio → no prosody features
  - Data not shared limits the potential for learning

Conversation technology will need to leverage advances in privacy-sensitive signal processing, learning, etc.

Other Ethical & Social Issues

- People should know when they are talking to a conversational agent – forget about the Turing test
- Architectures and learning strategies for language technologies should work for different languages
  - Note: prosody is used differently in different languages
- Systems should serve a diverse community of users, including age, dialect and cultural differences
  - Issue of bias in machine learning
  - User modeling should not reinforce stereotype bias
- Users can delete their data: limits reproducibility

Summary

- Spoken language ≠ noisy verbalized text
- In applying NLP to spoken language
  - Spoken language processing can benefit from a SpeechBERT
  - Multimodal integration of prosodic cues can benefit understanding
- For conversational systems
  - Need to track speaker and dialog state
  - User differences impact dialog policy & evaluation
- Spoken language processing research needs to be informed by and drive work in privacy and responsible ML