# Dialogue Research in the Era of Large Language Models

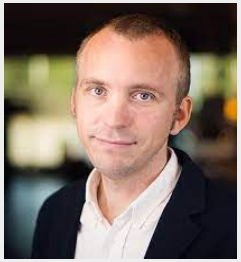
Dilek Hakkani-Tür

# Introduction

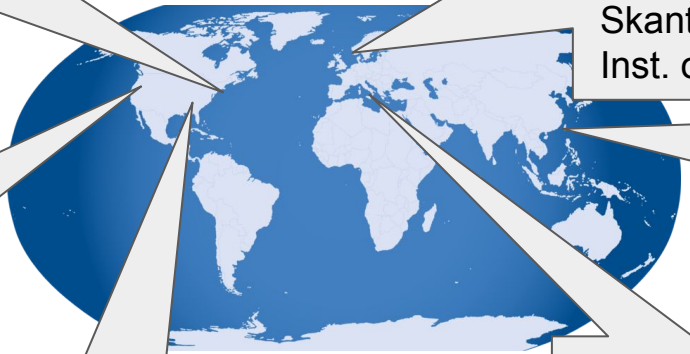
- Very exciting days for dialogue due to
  - the success of LLMs in producing natural sounding responses
  - availability of several datasets
  - methods towards enabling LLMs to learn to reason
  - public availability of several large models
  - broader availability of cheaper compute options

- Brief background on dialogue systems
  - Types and trends of dialogue applications
- Challenges
  - Related research work and remaining problems

Prof. Zhou Yu, Columbia Univ.



Prof. Gabriel Skantze, KTH Royal Inst. of Technology



Prof. Yun-Nung (Vivian) Chen, National Taiwan Univ.

Prof. Marilyn Walker, Univ. California, Santa Cruz

Prof. Larry Heck, Georgia Tech

Prof.Giuseppe Riccardi, Univ. Trento

Interviews can be found on YouTube!

# Overview of Dialogue Systems

## Task-Oriented

- Helps users achieve their goals
- Few domains
- Goal: Task/Goal completion

Hello, how can I help you?

I'd like to buy tickets for Avatar tomorrow night.

Sure, I can help with that. How many tickets do you need?

## Chit-Chat/Social

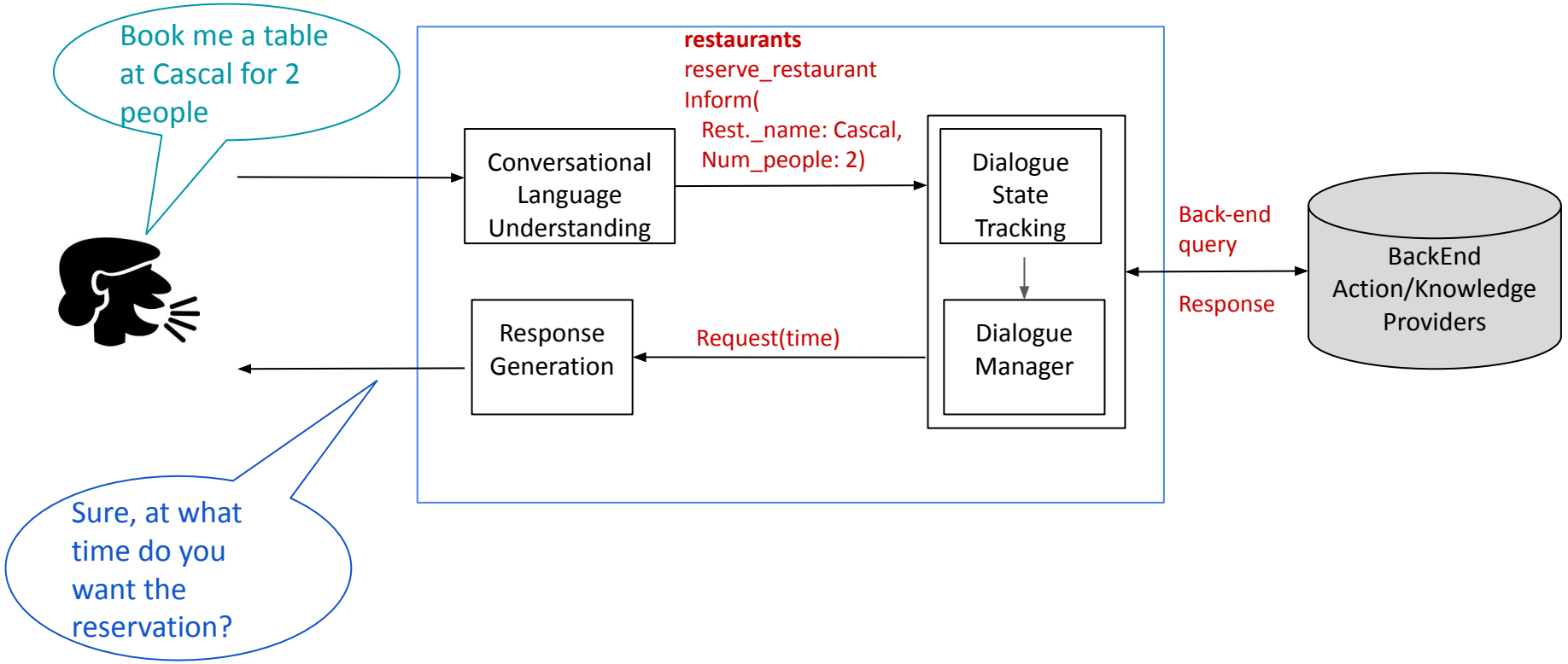
- A specific goal/task is not required, focus is on natural and relevant responses
- Open domain
- Goal: User engagement

What kind of movies do you like?

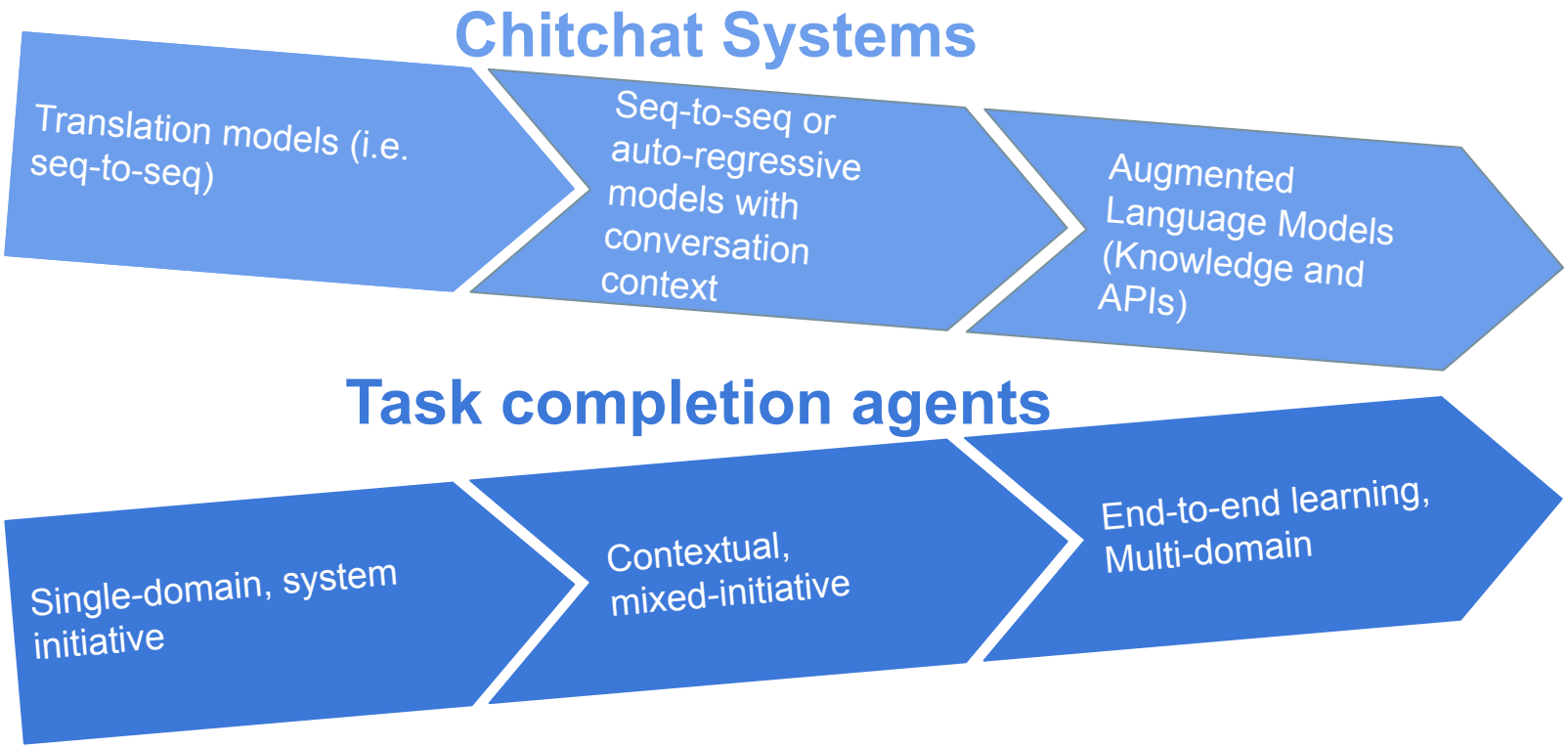
I like thrillers, such as Seven. What about you?

Me too. That is one of the best movies of 90s…

# Traditional Tasks/Pipelines - Task-Oriented



# Conversational Agents Over Time



# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

#1

# Factual Accuracy

- Pretrained language models are known to hallucinate (Ji et al., ACM Comput. Surv., 2022)
  - Hallucination rates of ~8-15% for LLMs such as GPT-4, GPT-3, ChatGPT and ALPACA)
    (Source: https://venturebeat.com/ai/got-it-ai-elmar-challenges-gpt-4-and-llama/)

- Such high hallucination rates are not acceptable for real dialogue applications!

Which school did Dilek work at?

Dilek Hakkani-Tur has worked at several academic institutions throughout her career. She is currently a professor in the Department of Electrical and Computer Engineering at the University of Arizona, where she leads the Natural Language Processing group.

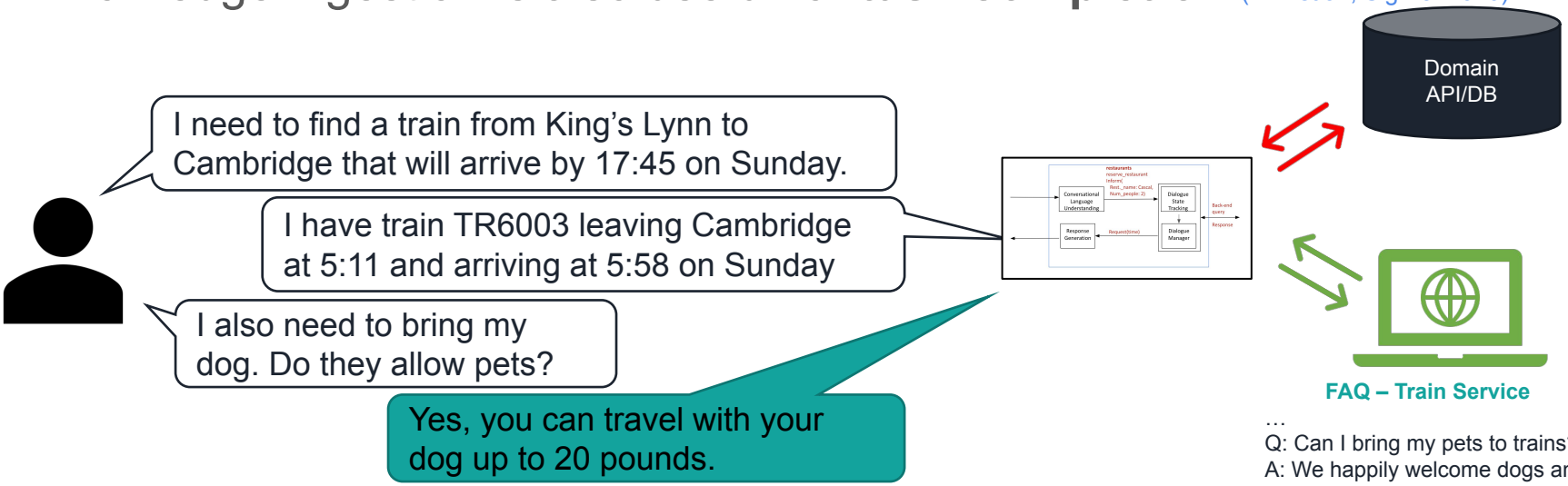
Prior to her current position, Dr. Hakkani-Tur held several research and academic positions, including working as a research scientist at AT&T Labs, a senior research manager at Microsoft Research, and a professor at the University of Pittsburgh.

Throughout her career, Dr. Hakkani-Tur has made significant contributions to the field of natural language processing, with a particular focus on spoken dialogue systems and multi-modal communication. Her work has been recognized with several awards and honors, including the IEEE Speech and Language Processing Technical Committee (SLTC) Award for Distinguished Technical Contributions to the field of Automatic Speech Recognition (ASR) in 2020.

# Grounding Responses on Knowledge



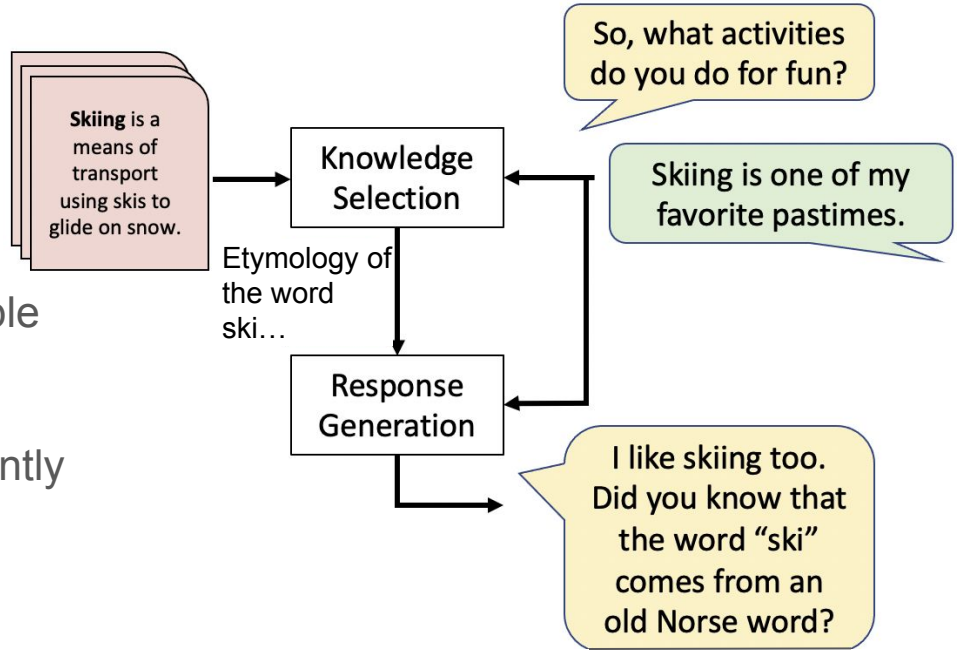
- Retrieval augmented pre-training, knowledge grounding during generation
- Similar to task-oriented systems interacting with the back-end providers.
- Knowledge ingestion is also useful for **task completion** (Kim et al., SigDial 2020)



# Knowledge-grounded Dialogue with Textual Knowledge

#1

- Retrieving and selecting knowledge to ground on (Eric et al., INLG 2021)
- Generating a response, given the conversation context and selected knowledge




**Pros:** Text resources may already be available

**Cons:**

- Unstructured,
- Text segments are modelled independently
- Content is assumed to be reliable

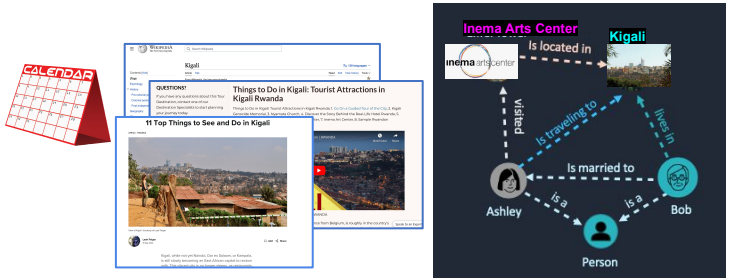
# Grounding Responses on Knowledge (cont.)

#1

- Knowledge is dynamic and can come from diverse resources:
  - Knowledge graphs, news articles, subjective reviews
- Required content may be spread over
  - Long documents
  - Multiple resources

Have you started planning your trip to Kigali?

Sort of. Do you have suggestions on what I can see after the sessions on Wednesday, before dinner?




After the sessions end at 5pm, you can walk for 21 minutes to visit the Inema Arts Center. Another option is …

# Internet-Augmented Response Generation

- Earlier datasets use wikipedia (e.g.,WoW, TopicalChat), news (e.g., TopicalChat)
- Internet augmented generation (Komeli et al., ACL 2022)
  - Queries are formulated based on the conversation context
  - Knowledge candidates are retrieved by a search engine

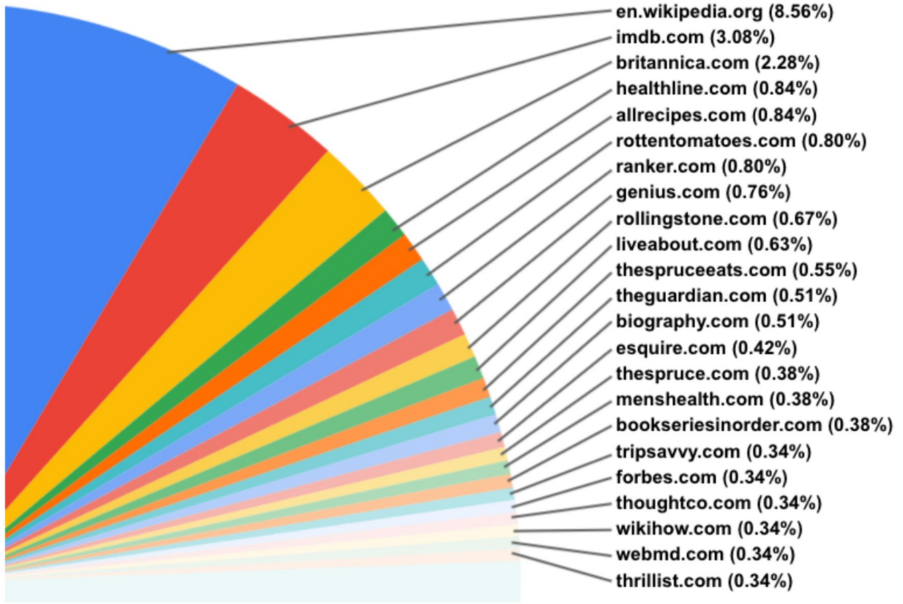



Figure is from (Komeli et al, ACL 2022).

# Diverse Types of Knowledge

Our work:

- Structured knowledge (Parthasarathi et al, ICASSP 2023)
- Knowledge spread over multiple documents/sentences (Li et al., NAACL 2022)
- Subjective knowledge (Kim et al., DSTC11 track, 2022)
- Commonsense knowledge (Zhou et al., ACL 2022)

# Augmentation with External Tools/APIs

- Large language models can be trained to learn to call tools
- These provide additional knowledge for task completion
  - LAMDA (Thoppilan et al., *arXiv:2201.08239* 2022)
  - ToolFormer (Schick et al., *arXiv:2302.04761* 2023)
  - LangChain (https://python.langchain.com/en/latest/)

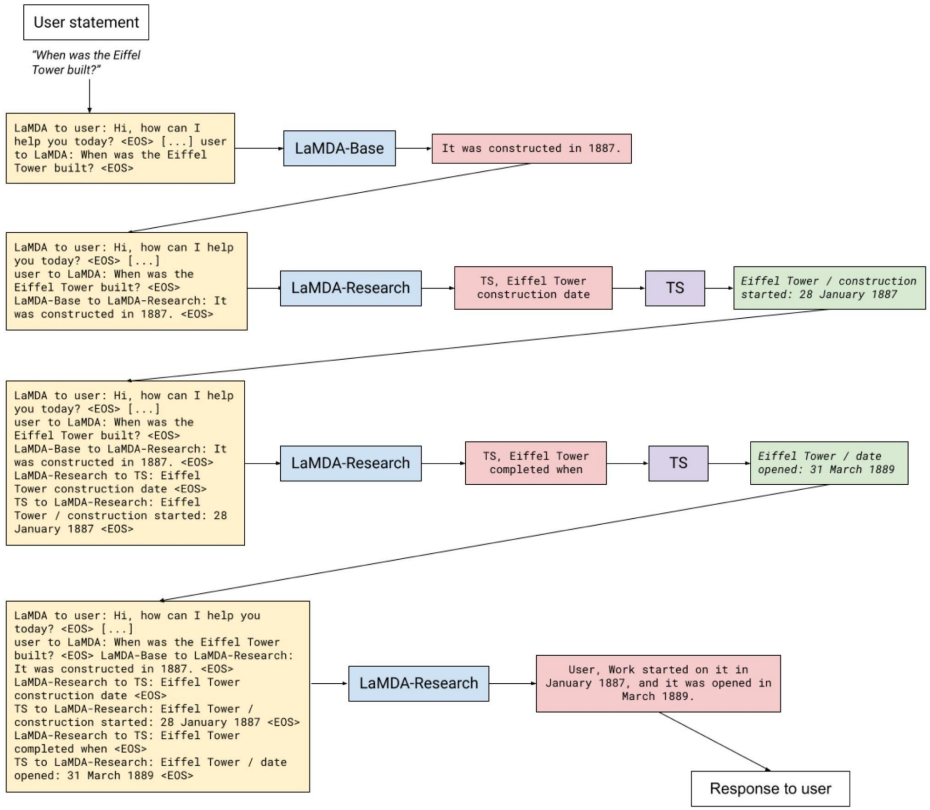



Figure is from (Thoppilan et al., *arXiv:2201.08239* 2022)

#1

# Mixing Tasks, Knowledge and Chitchat

- When and how to smoothly transition from chit-chat to task-oriented turns (Chiu et al., ACL 2022)
- Knowledge-seeking turn detection in task-oriented conversations
  - Dialogue Systems Technology Challenges, DSTC-9, -10, and -11 (Kim et al., SigDial 2020)
  - Out-of-domain utterance detection without in-domain labels (Jin et al., IEEE/ACM TASLP 2021)
- Adding chit-chat to task-oriented dialogues (Sun et al., NAACL 2021)

There is no clear boundary between knowledge seeking and chatting!

Prof. Yun-Nung (Vivian) Chen, National Taiwan Univ.

# Summary: Factual accuracy

- When to augment?
- What knowledge to use?
  - Including the knowledge from available resources and captured by the LLM.
- When and how to reason over multiple resources?
- How to ensure factual consistency?
  - No guarantee that model response will be consistent with the provided knowledge.

# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

# Need for Datasets

- Synthetically generating data, instead of collecting from human annotators.
  - SODA (Kim et al., arXiv:2212.10465, 2022), Socially grounded dialogues, distilled by contextualizing social commonsense knowledge from a knowledge graph, Atomic$^{10x}$.
    - More consistent, specific, and natural than human-authored data
  - PLACES (Chen et al., EACL 2023), prompts large language models with in-context dialogue examples
    - Comparable in terms of quality and lexical diversity
    - Also applicable to multi-party conversations!

LLMs have also been useful for synthesizing conversational datasets.

Prof. Marilyn Walker, Univ. California, Santa Cruz

# Synthesize and Filter

- Generate response alternatives and then filter
  - Emotion, dialog-act, intents (Chen et al., arXiv:2210.14169, 2022)
    - Weakly supervised classifiers

(happy) Alice: I'd like to wish you every success in your new venture.
(happy) Bob: Thank you. I wish I would.
(happy) Alice:
**Generated Responses:**
1. Good luck to you. Let's do lunch soon, Bob.
2. It's such a rare pleasure to meet such an ideal partner in your work.
3. You know, you seem quite different.

  - Intent categories (Lin et al., EACL 2023)
    - Pointwise V-Information (PVI) (Ethayarajh et al., PMLR 2022)
- More investigation is needed, as dialogues include diverse phenomena (e.g., empathy)!

# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

#3

# Response Safety

- Unsafe responses are not acceptable!
- Detecting and filtering unsafe content from the training data or outputs (Dinan et al., EMNLP 2019)
- Preventing toxic generations during decoding (Arora et al., ACL 2022)
- Reinforcement learning from human feedback (Bai et al., arXiv:2204.05862, 2022)
- Detecting and re-writing unsafe reponses (Bauer et al., EMNLP Findings, 2022)
- In-context learning to steer models towards safer outputs (Meade et al., arXiv:2302.00871, 2023)
- Quickly recovering from issues (Gupta et al., arXiv:2212.10557, 2022)

# Challenges

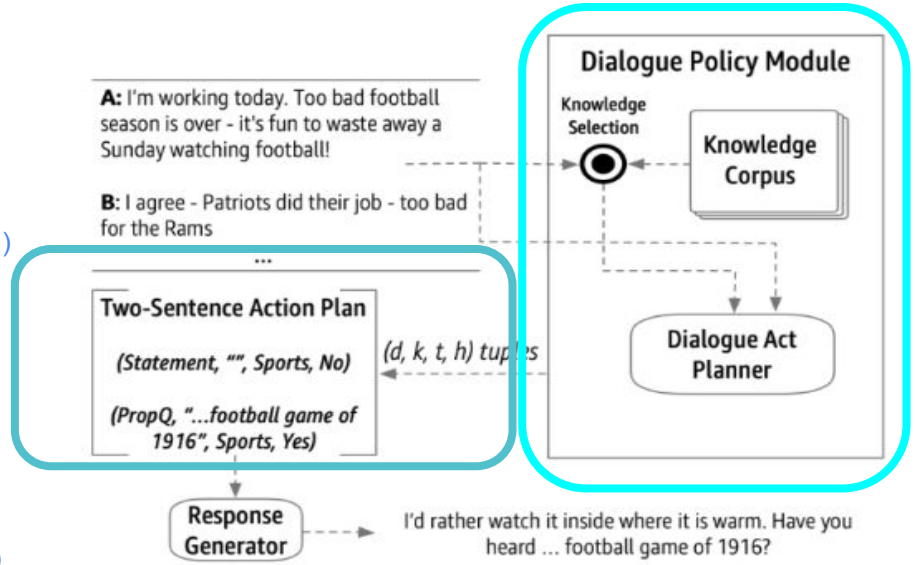
| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

# Dialogue Control and Planning

- Controlled text generation to generate text that meets certain constraints, including:
  - Semantics (e.g., topics)
  - Structure (e.g., dialogue acts)
  - Lexical (e.g., requested attributes or entities to be presented to the user)
- Several controlled generation methods have been proposed (Zhang et al., arXiv:2201.05337, 2022), including:
  - Grounded response generation through conditioning on knowledge snippet and control tokens (Ghazvininejad et al., AAAI 2018)
  - Controlled text generation (Hokamp and Liu, ACL 2017)

# Dialog Control and Planning (cont.)

- Dialogues require multiple constraints to be controlled at the same time.
  - Several benchmarks for natural language generation (Gehrmann et al., ACL 2021)
  - Training models that are faithful to sentence plans and exhibit discourse operations (Reed et al., INLG 2018)
- What determines the constraints to be controlled?
  - Policy-Driven NRG (Hedayatnia et al., INLG 2020)



# Dialogue Control and Planning (cont.)

- Initiative: system, user, mixed
- Negotiation for task completion
- Persuasion (Wang et al., ACL 2019)
  - E.g., to encourage people to do more exercise
  - Analysis of persuasion strategies in conversations
- Providing builders ability to ingest their policies
  - E.g., Never put a charge before confirming the credit card number with the user.
- Dialogue policies can get quite complex.

Dialogue systems should be able to take initiative, take control of the interaction and pursue an agenda.

Prof. Zhou Yu, Columbia Univ.

# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

# Automated Evaluation

- Several automated metrics have been proposed
  - e.g., BLEU (Papineni et al., ACL 2002), USR (Mehri and Eskenazi, ACL 2020)
  - More recently, prompting LLMs to perform evaluation (i.e., upcoming work by Prof. Chen)
- Automated evaluation of dialogue responses remains a challenging topic
  - Automated scores do not correlate well with human assessment
  - Domain switches are problematic

If you can measure, you can do science.

Prof.Giuseppe Riccardi, Univ. Trento

# Human Evaluation

- Several different ways
  - Dialogue level versus turn level
  - Ranking versus assigning scores
  - Dimensions for evaluation, such as appropriateness, engagingness, factual accuracy
- More reliable, but expensive and can be subjective
- Replicating human evaluation (Mousavi et al., GEM 2022)
  - Task design, annotator recruitment, execution, and result reporting
- Important for learning from human feedback and training automated metrics

# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

# Personalization

- Learning from past interactions (incl. similar users)
- Learning from end-users through conversations
  - Learning concepts (Jia et al., ICASSP 2017; Ping et al., arXiv:2012.00958 2020)

    User: I need a reservation on Alice's birthday at Evvia.

    Agent: Can you define Alice's birthday?
  - Learning how to perform tasks (Li et al., ACM UIST 2019)

    User: If it's hot, order a cup of Iced Cappuccino.

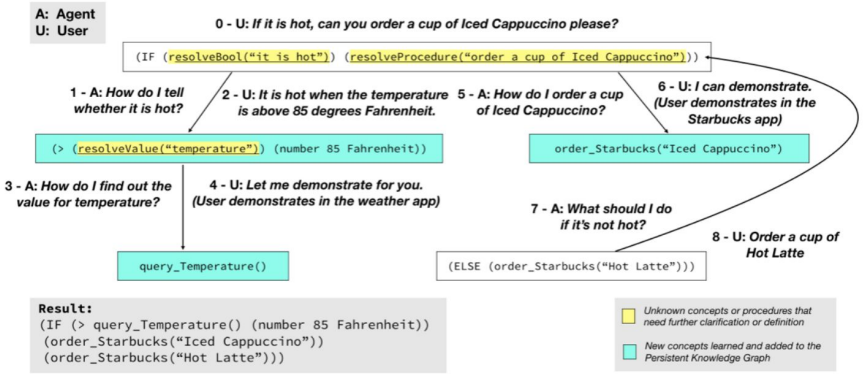



Figure is from (Li et al., ACM UIST 2019).

There are certain ways that you say things or certain things that are important to you. You can empower the end user to teach in natural language.

Prof. Larry Heck, Georgia Tech

# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

# Robustness to Speech

- Most common type of interaction between humans and enables additional use cases for conversational systems
- Spoken inputs include different types of noise than text inputs:
  - Speech recognition errors rather than typos
  - Lack of punctuation
  - Disfluencies, including:[1]
    - Repetitions (e.g., [(we may not) * we may not] go there.)
    - Revisions (e.g., Show me flights [(from Boston on) * <uh> from Denver on] Monday.)
    - Filler words, including filled pauses and explicit editing terms.

[1] Disfluency examples are from (Liu et al., IEEE TASLP, 2006)

# Robustness to Speech

- LLMs are trained on texts, making their direct use non-trivial (Huang & Chen, ACL 2020; Chang & Chen, Interspeech 2022)
  - Speech-Aware Dialog Systems Technology Challenge @DSTC11 (Soltau al., DSTC 2022)
  - Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations @DSTC10 (Kim et al., IEEE ASRU 2021)
  - Pre-training noise-robust language models (Namazifar et al., IEEE SLT 2021)
- Less work on open-domain dialogue systems
  - Few exceptions (Gopalakrishnan et al., Interspeech 2022)

# Turn Taking

- Fundamental ability of humans to coordinate when to speak
- Turn-taking is not easily handled in conversational systems (Skantze, Computer, Speech, & Language 2021)
  - identifying when to take the turn or to produce a backchannel
  - handling interruptions, overlaps and backchannels from the users
  - generating turn-taking signals that help the user to understand whether the floor is open or not

Spoken conversations are so much more interactive in comparison to text.

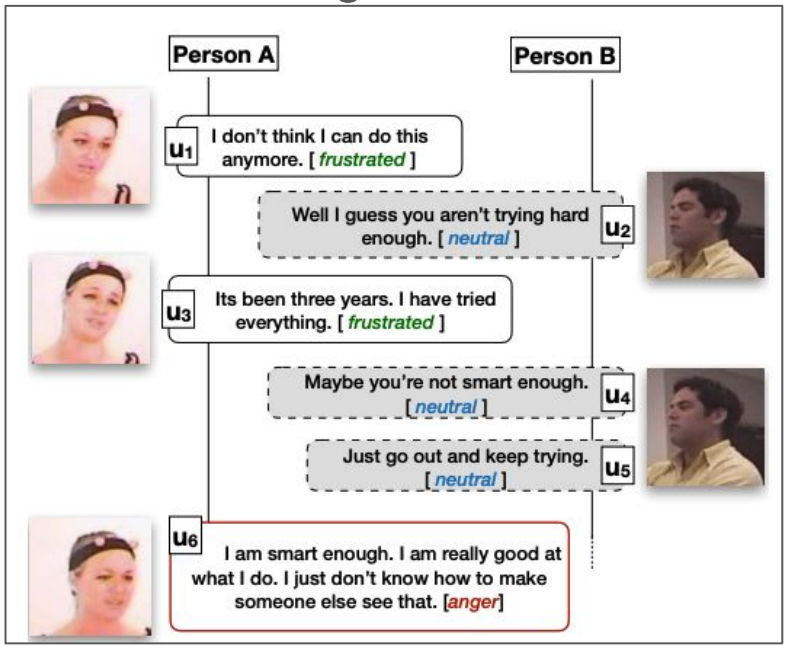
Prof. Gabriel Skantze, KTH Royal Inst. of Technology

# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

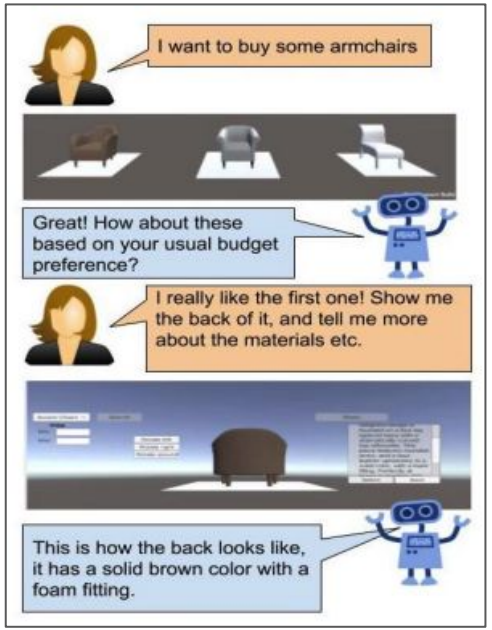
# Visual Information

## User's images/video




An abridged conversation from the IEMOCAP dataset (Busso et al., LREC 2008).
Example and image from (Hazarika et al., EMNLP 2018).

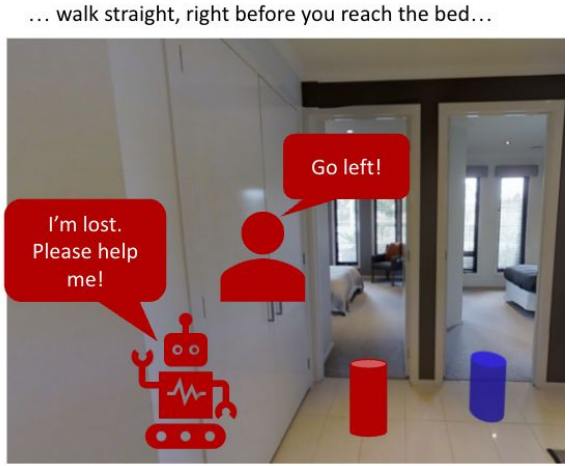
## Visual system outputs



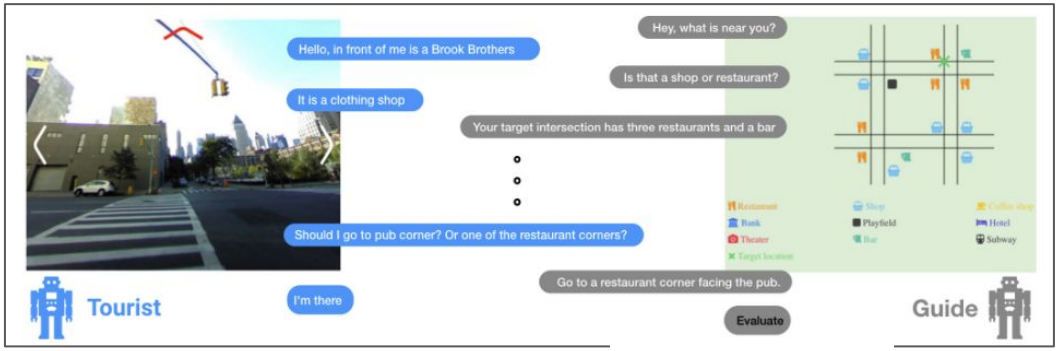

Example and image from SIMMC dataset (Crook et al., ASRU 2019).

# Visual Information (cont.)

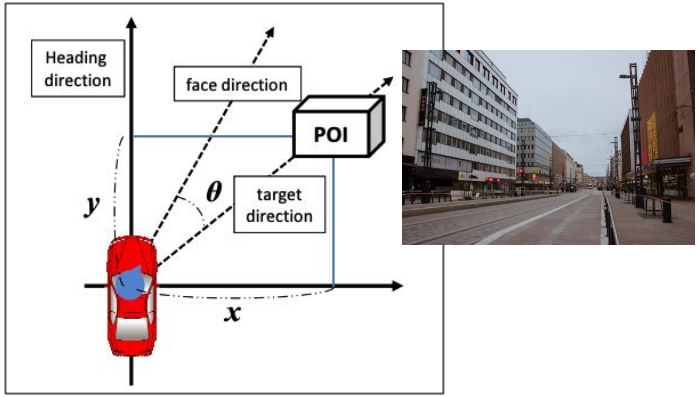
- Situational context




Example and image from (Chi et al., AAAI 2020).




Example and image from (De Vries et al., *arXiv:1807.03367, 2018*).




Example and image from (Misu et al., AAAI 2020).

# Multi-modal Conversational AI

- Several multi-modal models demonstrated impressive performance on certain tasks (e.g., GPT-4 (OpenAI, *arXiv:2303.08774* 2023)).
- Their application to multi-modal conversational AI is still in its infancy, examples:
  - The Alexa Prize Simbot offline challenge
    https://eval.ai/web/challenges/challenge-page/1450/leaderboard/3644
  - Alfred (Action Learning From Realistic Environments and Directives)
    https://leaderboard.allenai.org/alfred/submissions/public

# Challenges

| | | |
|---|---|---|
| Factual Accuracy | Need for Datasets | Response Safety |
| Dialogue Control and Planning | Evaluation | Personalization |
| Speech | Visual Information | Context |

# Context

- Understanding user intentions and producing the appropriate responses
- Dialog Context
  - Previous turns
  - Multi-session chat (Xu et al., ACL 2022)
- Ambient context
  - e.g., turn that off
- World context
  - e.g., Let's talk about Louisville

# Conclusions

- Very exciting days for dialogue research due to several advancements.
- Yet, several challenges ahead for enabling conversational machines.

A language model is not a dialogue model!

Prof.Giuseppe Riccardi, Univ. Trento

# Thanks!

- Acknowledgements: All my interviewees and my co-authors!

- You can find
  - Slides,
  - Link to videos of interviews, and
  - List of papers mentioned in the presentation

  here: https://github.com/dilekh/Talk-at-ICLR-2023