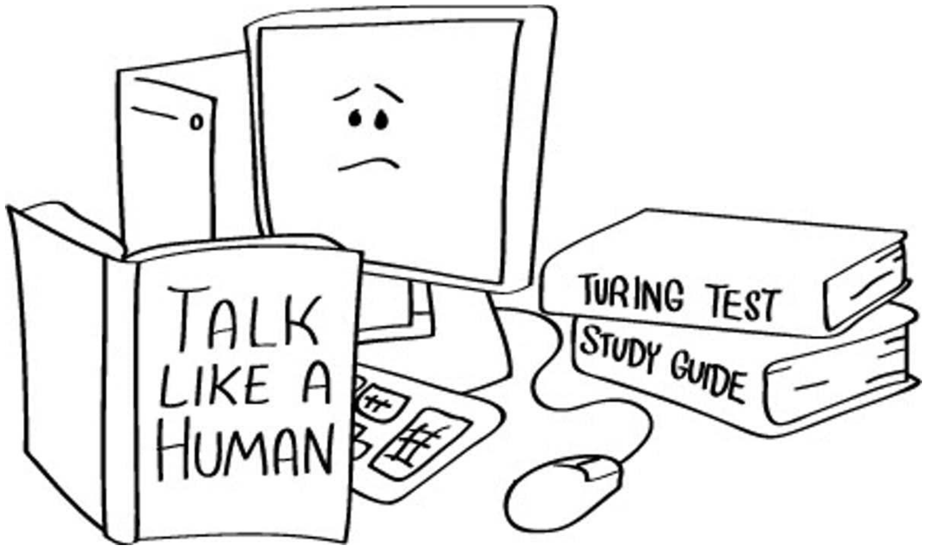
10 things you should know about dialogue

Milica Gašić,
Dialog Systems and Machine Learning
Heinrich Heine University Düsseldorf

Dialogue as a core AI problem

- Turing poses dialogue as a core AI problem (*Turing test*)
- Dialogue is hard: infinite possible trajectories of system and user turns
- We can always think of a dialogue that was never produced before
- Dialogue is an AI complete problem



Turing, Computing Machinery and Intelligence, 1950

Shapiro, Encyclopedia of Artificial Intelligence, 1992

1. What is the point?

Heinrich Heine Universität Düsseldorf

Task-oriented vs chat-based

- Humans do not make a strict distinction between task oriented and chat dialogue, while modelling approaches do
- Task oriented dialogues
  - typically have a goal or a number of goals
  - have well-defined scope of conversation
- Social dialogue approaches
  - typically do not model a goal
  - allow the conversation to span over a huge number of topics and impose no restriction on the vocabulary
  - model emotion and sentiment
- Today the approaches are intertwined [and that is a good thing!]

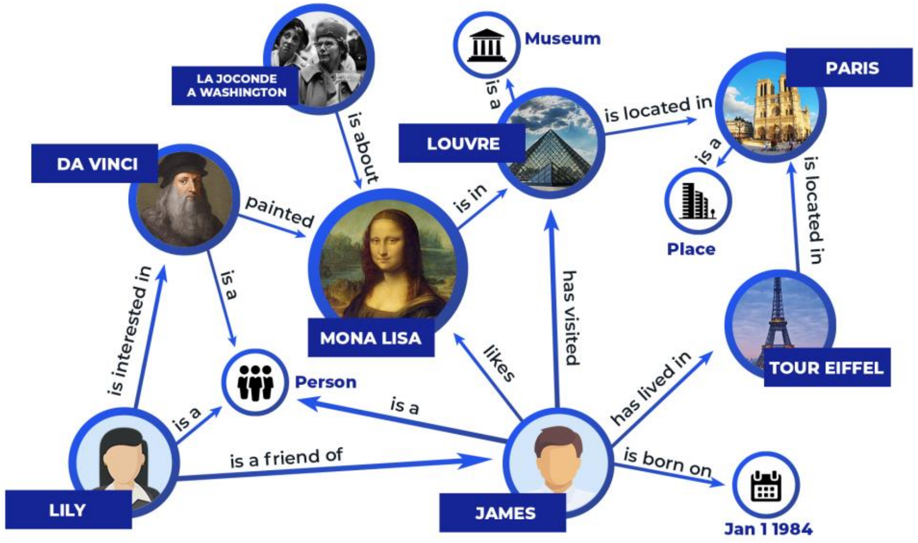
Zue et al, JUPITER: a telephone-based conversational interface for weather information, 2000

Feng et al, Sounding board – University of Washington's Alexa Prize submission, 2017

1. What is the point?

The concept of ontology

- Task-oriented dialogues typically interface a data-base
- The database is described by an underlying ontology
- Simplest ontology:
  - domain, slots, value
- Can be more complex, eg knowledge graph



1. What is the point?

Heinrich Heine
Universität Düsseldorf

The concept of dialogue act

- Dialogue act formalism describes meaning encoded in each dialogue turn
  - Relation to ontology
  - Intention of the speaker
  - Relation to logic
  - Context
  - Partial information from speech recognizer (primitive dialogue acts)
- Today, with the advent of NN approaches, the intermediate dialogue act formalism is disappearing

Traum, 20 questions on dialogue act taxonomies, 2000

2. When to speak?

The concept of dialogue turn

- Dialogue can be described in terms of system and user turns
  - System: How may I help you?
  - User: I’m looking for a restaurant
  - System: What kind of food would you like?
  - …
- Turn taking can be more complex and characterised by barge-ins
  - System: How may I... User: I’m looking for a restaurant
- Back channels
  - User: I’m looking for a restaurant [System: uhuh] in the centre of town

Skantze and Schlangen, Incremental Dialogue Processing in a Micro-Domain, EACL, 2009

Paetzel et al, “So, which one is it?” The effect of alternative incremental architectures in a high-performance game-playing agent, 2015

3. Context, context, context

Heinrich Heine
Universität Düsseldorf

Dialogue state

- Understand the user
- Respond to the user
- Conduct the conversation beyond question answering
- There are infinite plausible dialogue trajectories
- The dialogue state summarises what is important in the dialogue so far
  - Dialogue history
  - User goal
  - Grounding information
  - Co-reference resolution

Clark and Brennan, Grounding in Communication, chapter 7. APA, 1991.

Larson and Traum, Information state and dialogue management in the TRINDI dialogue move engine toolkit, Natural Language Engineering, 5(3/4):323–340, 2000.

3. Context, context, context

Dialogue state tracking

- Maintaining dialogue state throughout dialogue is essential
- Supervised learning task
- Approaches
  - Bayesian networks
  - Neural networks
- Things to consider:
  - How well does the tracking perform on its own?
  - Can it run real time?
  - Does it support user changing their mind?
  - What happens when you introduce new values to the ontology?

3. Context, context, context

Heinrich Heine Universität Düsseldorf

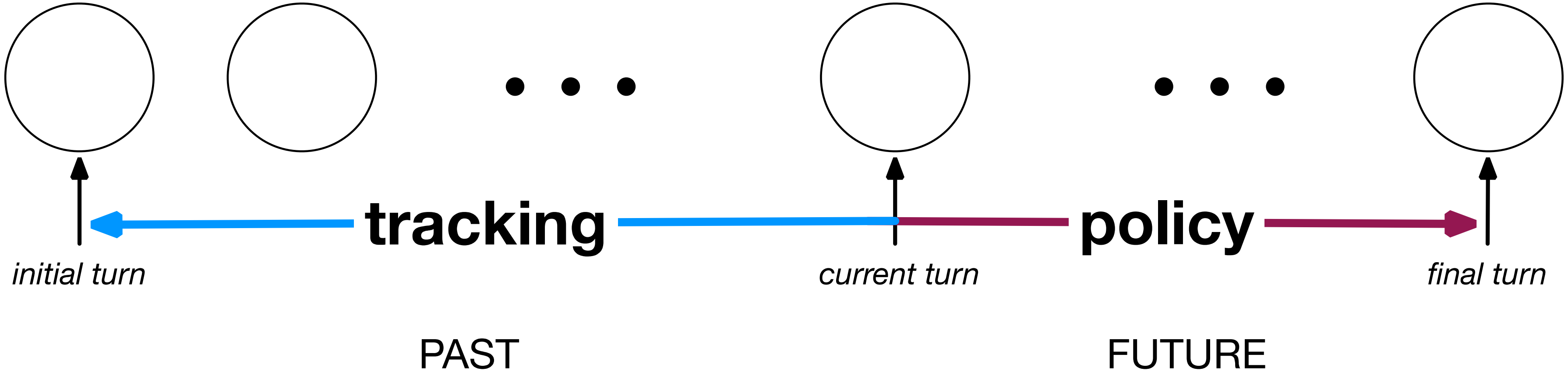
TripPy – value independent dialogue state tracker

- Dialogue state is constructed using span prediction
- TripPy deploys a triple copy mechanism:
  1. Span prediction may extract values directly from the user input;
  2. a value may be copied from a system inform memory that keeps track of the system's inform operations;
  3. a value may be copied over from a different slot that is already contained in the dialog state to resolve coreferences within and across domains.

Heck et al, <u>TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking</u>, SIGDIAL, 2020

4. A game to play

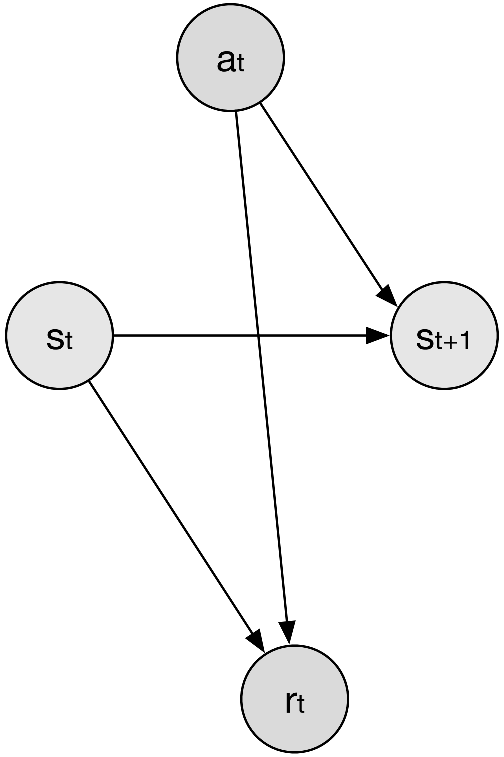
Planning



4. A game to play

Dialogue as a Markov decision process

- Dialogue can be defined in terms of dialogue states, system actions (responses) and associated rewards
- Markov decision process postulates that the next state depends only on the previous state and the action
- Reinforcement learning is an attractive framework for optimising dialogue policy
- Dialogue policy
  - decides which action to take in a given dialogue state
  - steers dialogue towards goal completion

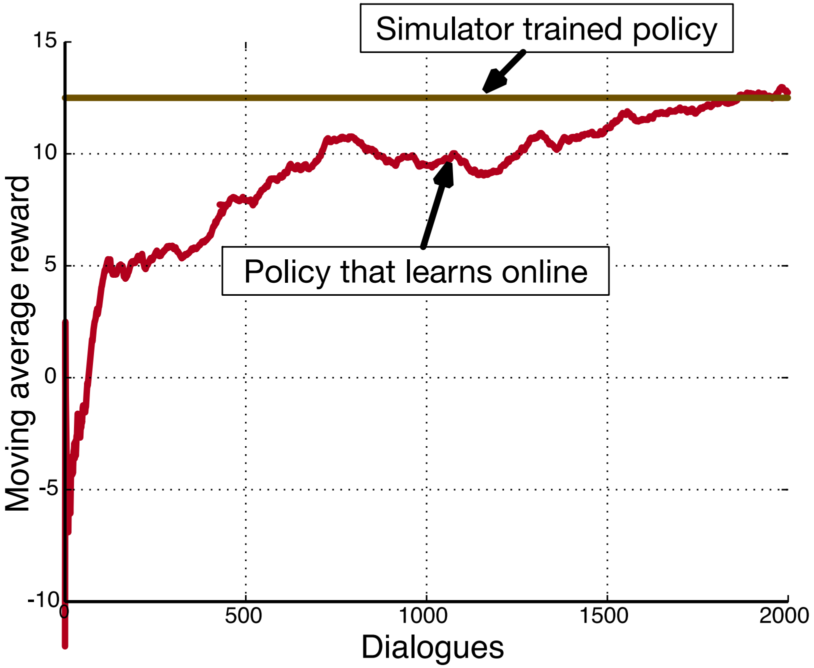


Levin et al, <u>A Stochastic Model of Human-Machine Interaction for Learning Dialogue Strategies</u>, Eurospeech, 2000

4. A game to play

Learning from human interaction

- Dialogue policy must efficiently explore possible actions



Gašić et al, <u>On-line policy optimisation of spoken dialogue systems via live interaction with human subjects</u>, ASRU, 2011

5. Are you sure?

Heinrich Heine
Universität Düsseldorf

State of the art dialogue state trackers

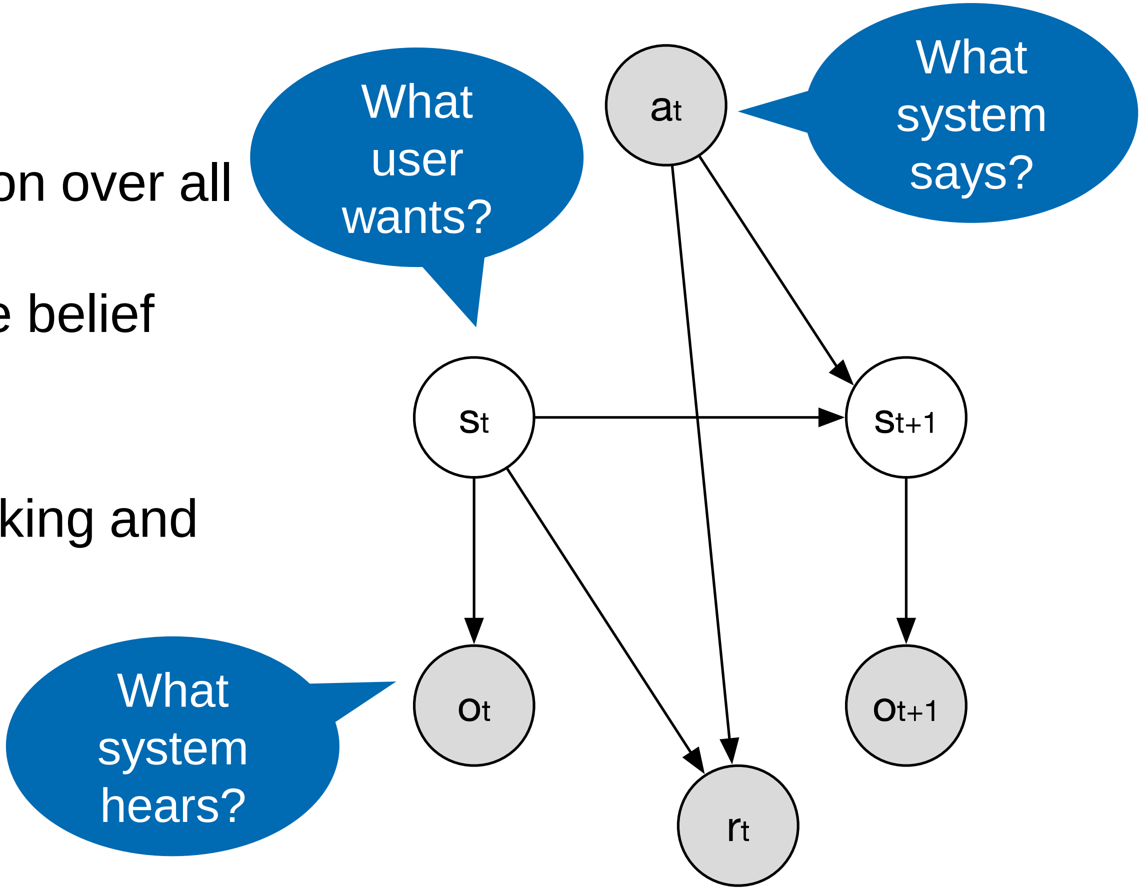
	MultiWOZ 2.0		MultiWOZ 2.1	
Model	Joint Accuracy	Slot	Joint Accuracy	Slot
MDBT (Ramadan et al., 2018)	15.57	89.53		
GLAD (Zhong et al., 2018)	35.57	95.44		
GCE (Nouri and Hosseini-Asl, 2018)	36.27	98.42		
Neural Reading (Gao et al, 2019)	41.10			
HyST (Goel et al, 2019)	44.24			
SUMBT (Lee et al, 2019)	46.65	96.44		
TRADE (Wu et al, 2019)	48.62	96.92	45.60	
COMER (Ren et al, 2019)	48.79			
DSTQA (Zhou et al, 2019)	51.44	97.24	51.17	97.21
DST-Picklist (Zhang et al, 2019)			53.3	
SST (Chen et al. 2020)			55.23	
TripPy (Heck et al. 2020)			55.3	
SimpleTOD (Hosseini-Asl et al. 2020)			55.72	

5. Are you sure?

Heinrich Heine Universität Düsseldorf

Modelling uncertainty

- In each dialogue turn we maintain the distribution over all possible states – belief state
- Instead of tracking dialogue states, we track the belief states
- The computational complexity explodes
- This creates difficulties both for belief state tracking and policy optimisation



Young et al, Pomdp-based statistical spoken dialog systems: A review, IEEE, 2013

Modular vs end-to-end systems

- Traditional dialogue systems are a pipeline of modules
  - Automatic Speech Recognition
  - Natural Language Understanding / Dialogue State Tracking
  - Policy
  - Natural Language Generation
  - Text-to-Speech Synthesis
- If we view the human brain as a giant neural network it is reasonable to think that we might produce an artificial neural network which takes words as input and outputs words
- There are many problems, one being the difficulty to incorporate planning.

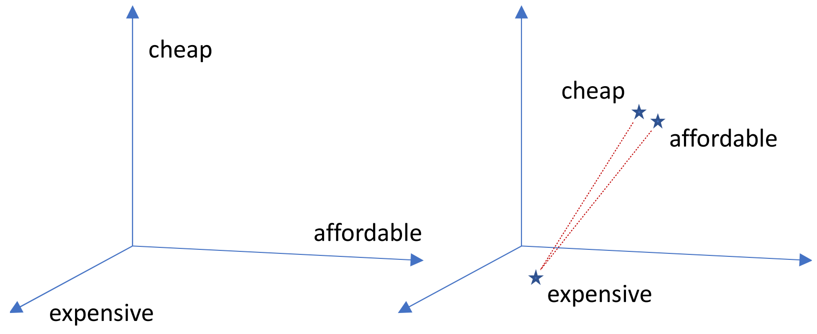
Zhao et al, Rethinking Action Spaces for Reinforcement Learning in End-to-end Dialog Agents with Latent Variable Models, NAACL, 2019

7. “The meaning of a word lies in its use”

Heinrich Heine Universität Düsseldorf

Symbolic vs distributed representations

- When underlying dialogue operation is described in terms of symbols for domains, slots and values
  - we need labelled training data
  - we need to perform delexicalization
  - we cannot associate words unseen in data with symbols
- Distributed representations (aka word vector embeddings)
  - utilise large unlabelled corpora
  - provide semantic similarity between words
  - remove the need for delexicalization
  - have better generalisation capabilities



Mrksic et al, Counter-fitting word vectors to linguistic constraints, NAACL, 2016

8. Am I doing well?

Heinrich Heine Universität Düsseldorf

Metrics

- Automatic metrics (BLEU, ROUGE, METEOR) are appealing because of their simplicity but are often misleading
- Success or completion rates measure how well the system can fulfil the user goal and are indispensable for task-oriented dialogue
- User satisfaction is very important but very difficult to measure
- We need to take into account efficiency measures (#dialogue turns, response time)
- Additional measures to consider:
  - Naturalness
  - Informativeness
  - Fluency
  - Readability (fluency in context)

Walker et al, PARADISE: A Framework for Evaluating Spoken Dialogue Agents, ACL, 1997

Stent et al, Evaluating Evaluation Methods for Generation in the Presence of Variation, CICLing, 2005

9. I am talking to you

Human-in-the loop

- User-centric technology – eventually we need to evaluate with humans
- Typical setting for evaluation:
  - Recruit volunteers
  - Produce tasks for them (eg: book a 3 star hotel in the centre of town)
  - Let them talk to the system
- In-lab testing is very laborious; can only collect a small number of dialogues
- One way to scale up is via crowdsourcing

9. I am talking to you

hhu
Heinrich Heine
Universität Düsseldorf

Let’s Go! Challenge

- Ravenclaw dialogue system was connected to Pittsburgh bus information phone line after working hours
- People would call the system and ask (eg “When is the next 61C arriving?”)
- These dialogues were provided as training data for the systems in the challenge
- Best performing systems were connected to the phone line
- Very interesting and unexpected outcomes
- IMPORTANT: This is a REAL USER experiment

Bohus and Rudnicky, The RavenClaw Dialog Management Framework: Architecture and Systems, CSL 2009

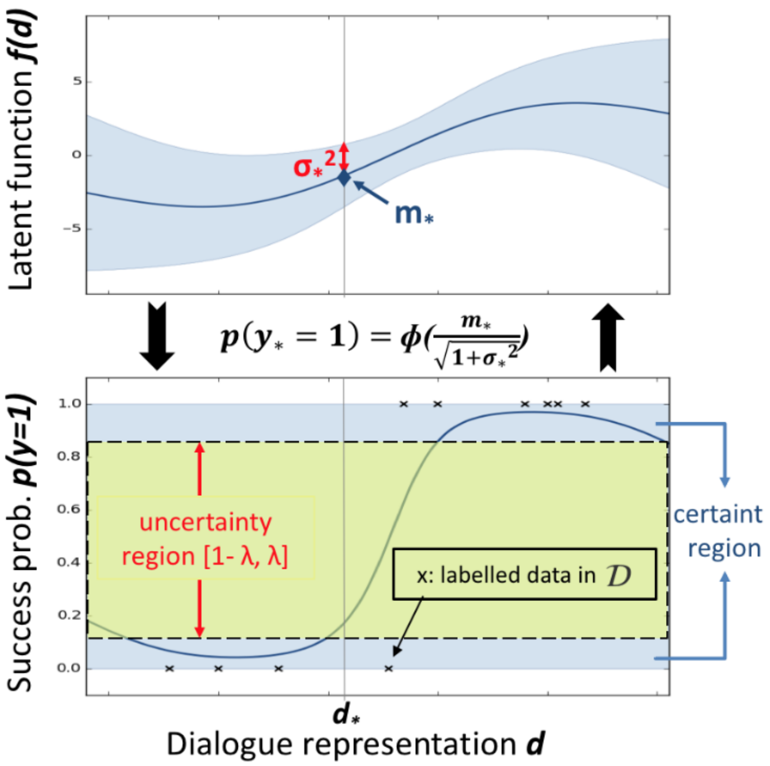
Raux et al, Let's Go Public! Taking a Spoken Dialog System to the Real World, INTERSPEECH, 2005

Black et al, Spoken dialogue challenge 2010, SLT 2010

9. I am talking to you

Heinrich Heine
Universität Düsseldorf

Dealing with unreliable input from the users



Su et al, On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems, ACL, 2013

10. No data like more data

Training and testing corpora

- Large corpora for chit-chat
- Small corpora for task-oriented dialogues (~2K)
- For a multi-domain set-up we need substantially more data
- Wizard-of-Oz set-up is one way of collecting more dialogues

10. No data like more data

MultiWOZ dataset

Metric	DSTC2	SFX	WOZ2.0	FRAMES	KVRET	M2M	MultiWOZ
# Dialogues	1,612	1,006	600	1,369	2,425	1,500	**8,438**
Total # turns	23,354	12,396	4,472	19,986	12,732	14,796	**115,424**
Total # tokens	199,431	108,975	50,264	251,867	102,077	121,977	**1,520,970**
Avg. turns per dialogue	14.49	12.32	7.45	**14.60**	5.25	9.86	13.68
Avg. tokens per turn	8.54	8.79	11.24	12.60	8.02	8.24	**13.18**
Total unique tokens	986	1,473	2,142	12,043	2,842	1,008	**24,071**
# Slots	8	14	4	**61**	13	14	25
# Values	212	1847	99	3871	1363	138	**4510**

Budzianowski et al, MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, EMNLP, 2018

Conclusion

- Dialogue requires much more sophistication than eg a seq2seq model provides
- Important lessons to be drawn from previous approaches
- Deep learning models which draw from these lessons achieve state of the art results
- Still there is a lot more we need to achieve
  - Current state tracking approaches are wrong almost every second turn
  - The available labelled data sets are still very small given the difficulty of the problem
  - We are building user-centric technology and evaluating on measures such as BLEU
  - Reinforcement learning is promising but difficult in an end-to-end setting
  - …