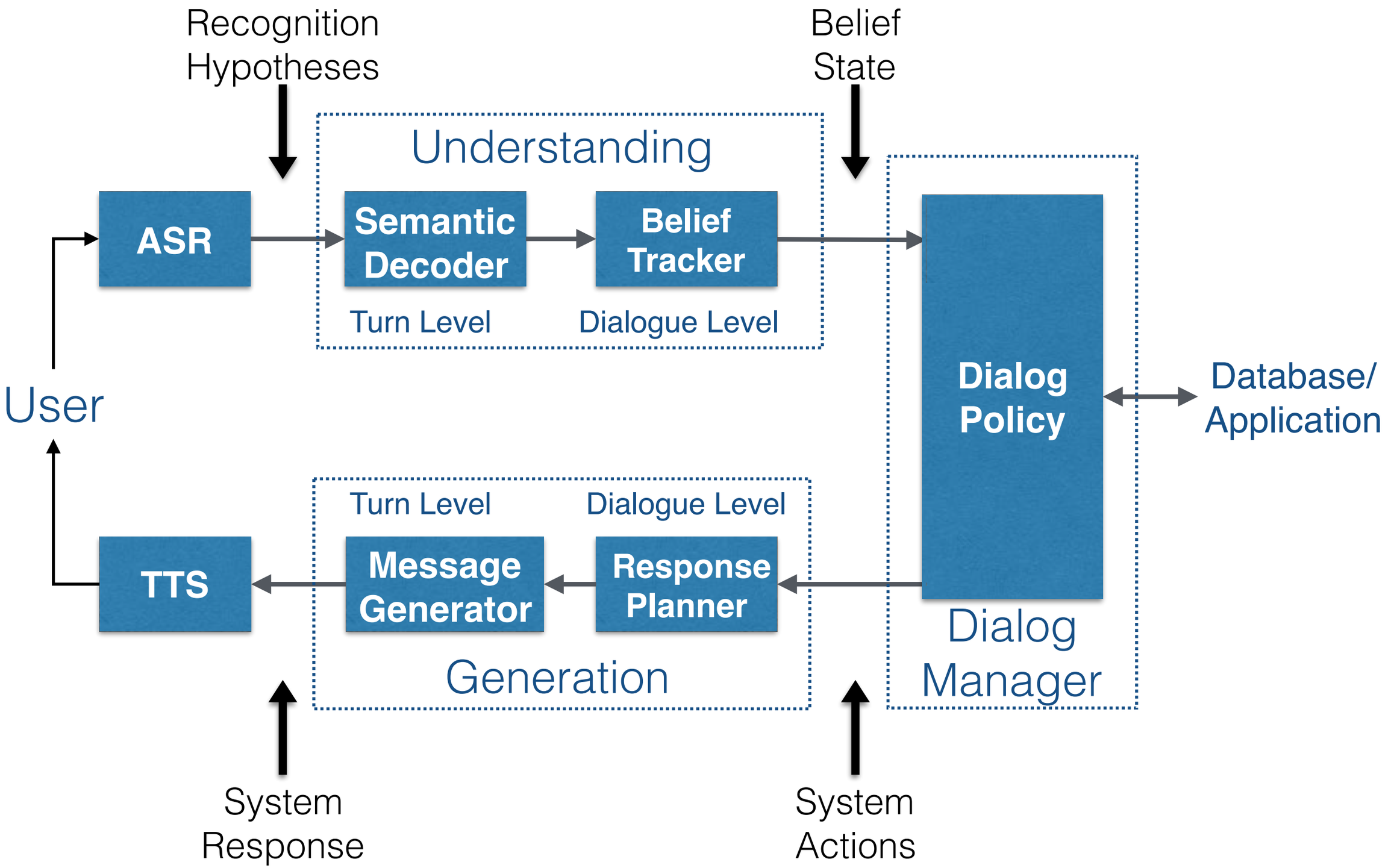
Statistical Spoken Dialogue Systems and the Challenges for Machine Learning

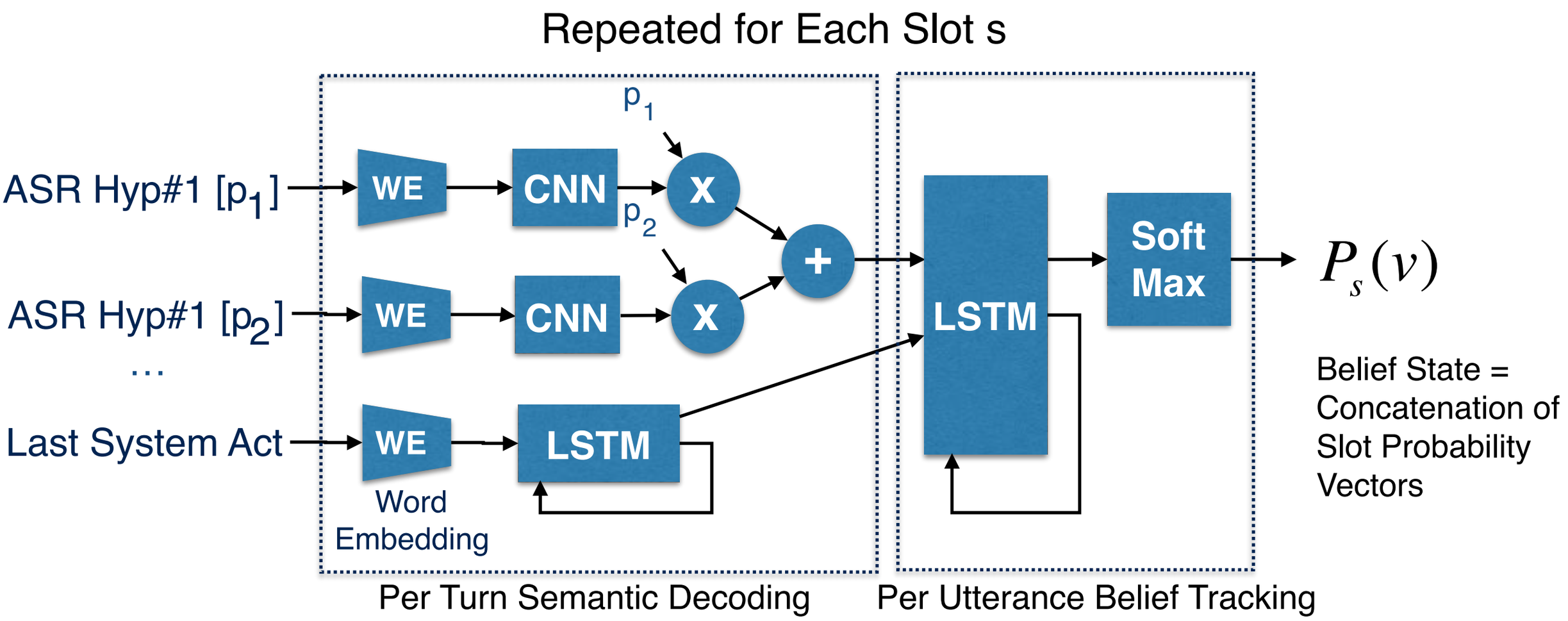
Steve Young

Dialogue Systems Group
Machine Intelligence Laboratory
Cambridge University Engineering Department
Cambridge, UK

Dialog System Architecture

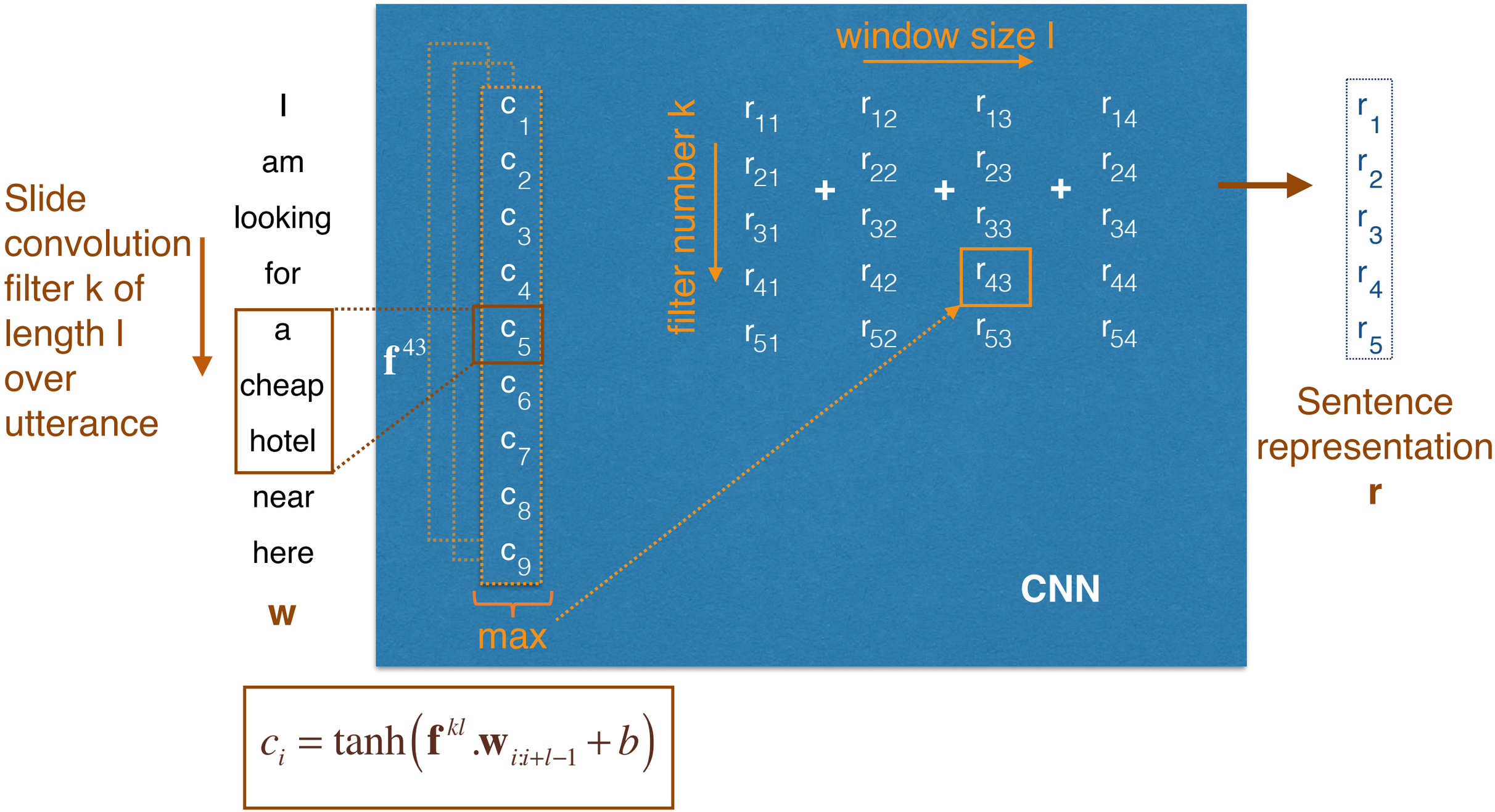


Understanding: ASR -> Beliefs



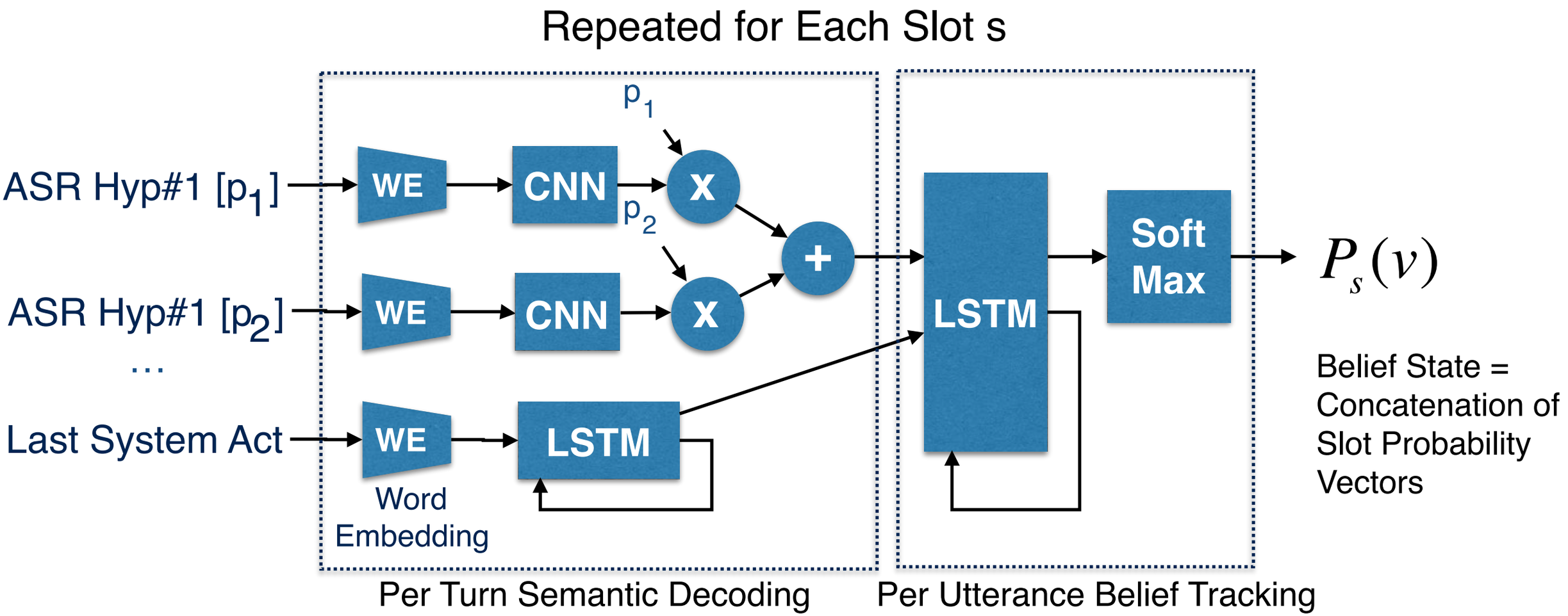
Using a CNN to Extract Lexical Features

CNN is the key component: it scans each utterance applying convolution windows of 1, 2, 3, 4, … words



$$c_i = \tanh\left(\mathbf{f}^{kl}.\mathbf{w}_{i:i+l-1} + b\right)$$

Understanding: ASR -> Beliefs



Henderson, M., et al. (2014). Word-Based Dialog State Tracking with Recurrent Neural Networks. SigDial 2014, Philadelphia, PA.

Rojas-Barahona, L., et al. (2016). Exploiting Sentence and Context Representations in Deep Neural Models for Spoken Language Understanding. Coling, Osaka, Japan.

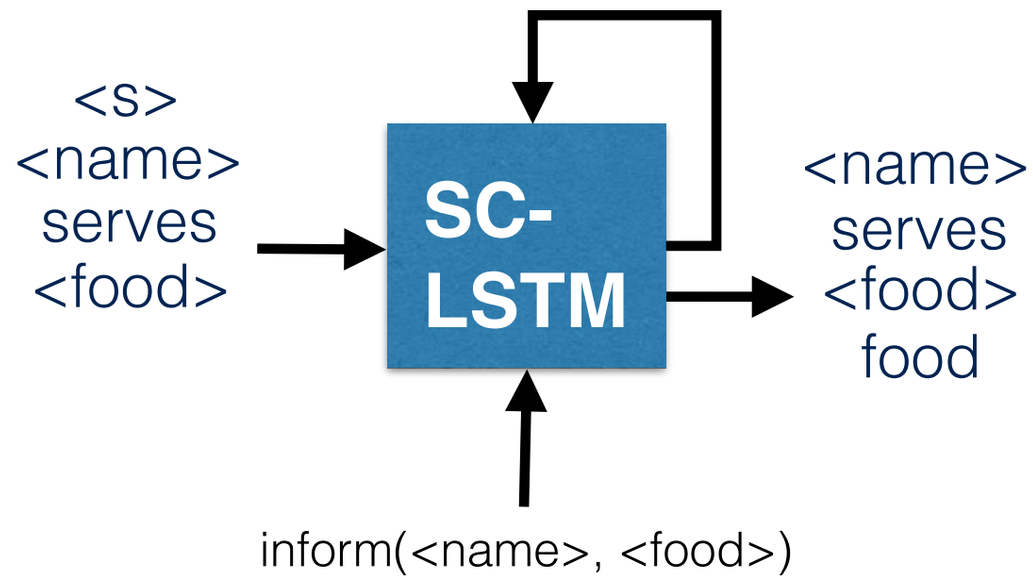
Mrksic, N., et al. (2016) Neural Belief Tracker: Data-Driven Dialogue State Tracking. arXiv:1606.03777

Generation: actions -> words

Need to convert abstract system actions to natural language e.g.

inform(name="The Peking", food="chinese") → "The Peking serves chinese food"

training



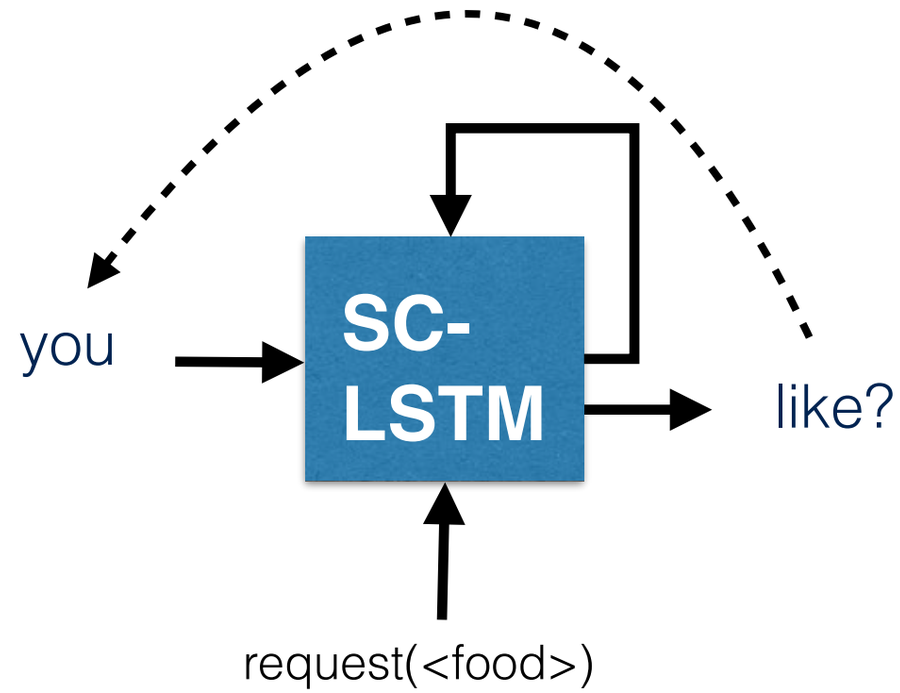
Generation: actions -> words

Need to convert abstract system actions to natural language e.g.

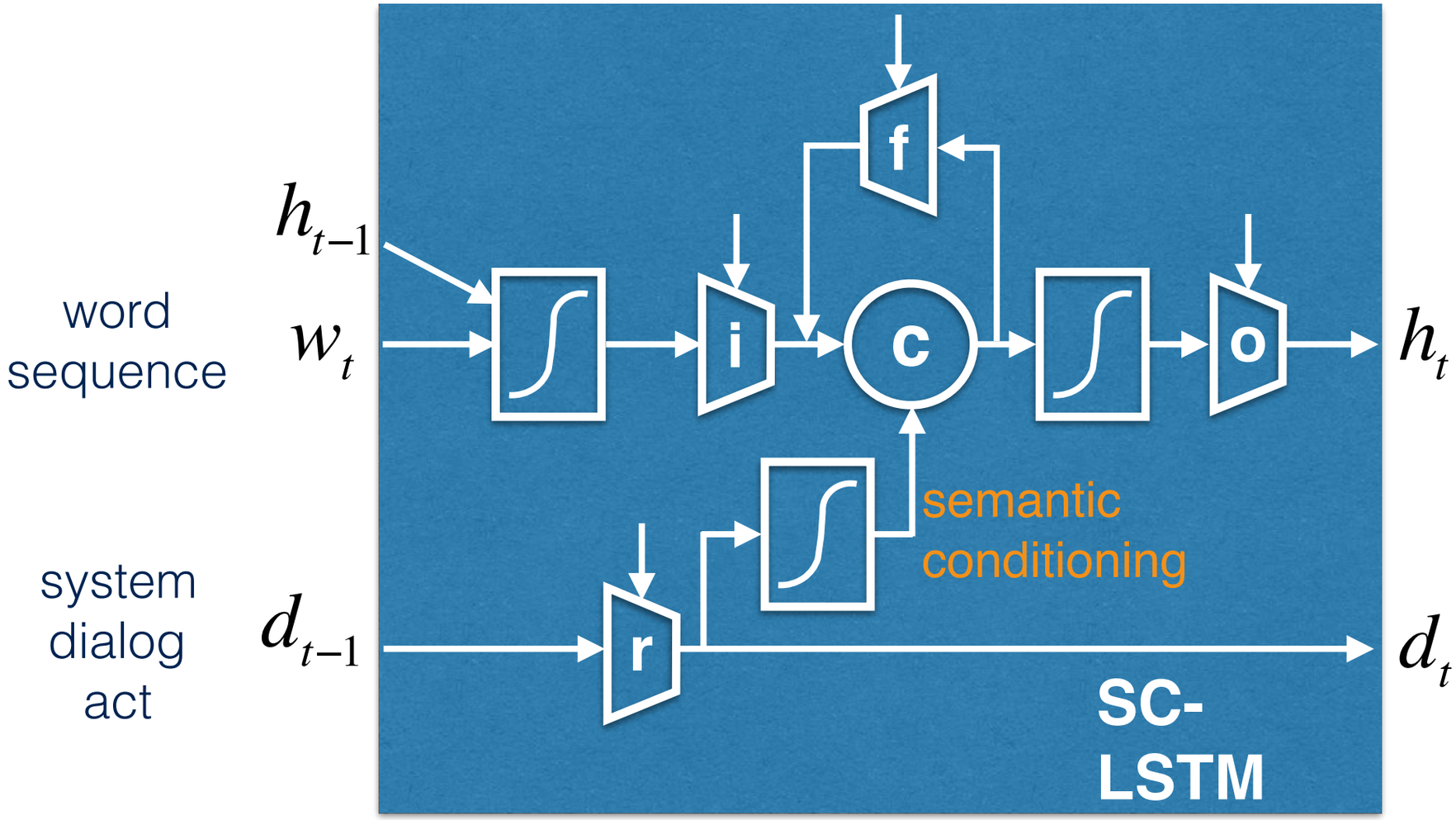
inform(name=<name>, food=<food>) → “ <name> serves <food> food”

Solution: delexicalise the training data, and train a conditional LSTM

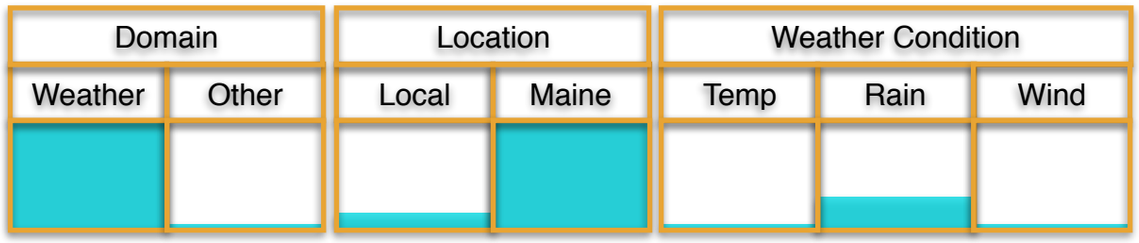
running



Semantically constrained LSTM



Dialog Manager



$$\mathbf{b} \xrightarrow{\pi} a$$

Actions: request, confirm, inform, execute, etc

1. Belief state $\mathbf{b}$ encodes the state of the dialog, including all relevant history.
2. Belief state is updated every turn of the dialog.
3. The policy π determines the best action to make at each turn via a mapping from the belief state $\mathbf{b}$ to actions a.
4. Every dialog ends with a reward: +ve for success, -ve for failure. Plus a weak -ve reward for every turn to encourage brevity.
5. Reinforcement Learning is used to find the best policy.

Reinforcement Learning

Policy: $\pi(\mathbf{b},a):\mathbb{R}^n \times A \rightarrow [0,1]$

Reward: $R = \sum_{\tau=1}^{T} r(\mathbf{b}_\tau, a_\tau)$ NB: no discounting:

Problem: find $\pi^* = \arg\max_{\pi} \{E[R \mid \pi]\}$

Policy Representation

$$\pi(\mathbf{b},a):\mathbb{R}^n \times A \rightarrow [0,1]$$

n ~ 20 - 100

|A| ~ 200+

- Gaussian Processes: data efficient, includes explicit confidence on Q-value. Can support large n, but action space |A| limited.
- Deep Neural Networks: scale well on both n and |A|, but no built-in confidence measure and poor convergence properties.

Training Data

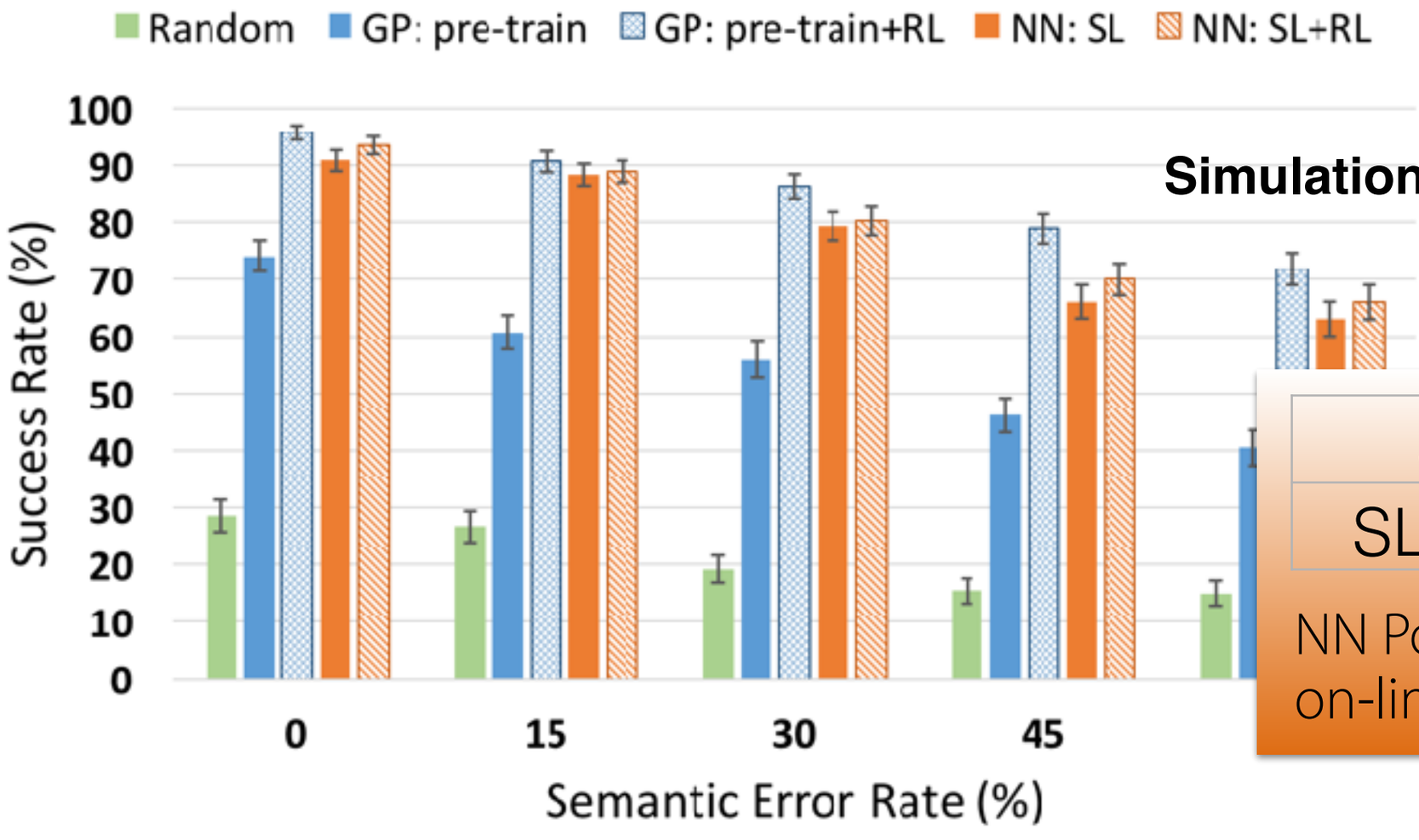
- Ideally, train directly on interactions with real users but
  - ✦ training even a small domain may require around 5k dialogues (many in exploration mode)
  - ✦ reward signal is hard to measure (see later)
- In practice, train in stages
  - ✦ initialise with corpus data
  - ✦ train/test on user simulator
  - ✦ tune on real users

Optimisation Algorithms

- Policy Iteration
  - ✦ GP Sarsa
  - ✦ Deep Q-learning
- Policy Gradient
  - ✦ Natural Actor Critic
- “Black box” methods
  - ✦ Trust Regions

NAC trained Neural Net Policy vs GP Policy

1. NN policy: 1 common 32 node tanh hidden layer. Action outputs encoded via 2 softmax output partitions and 6 sigmoid partitions
2. Pre-trained (using SL for NN and prior for GP) on 720 dialogs from Cambridge restaurant domain.
3. Optimised (using RL) on 5000 simulated dialogues



Simulation Results

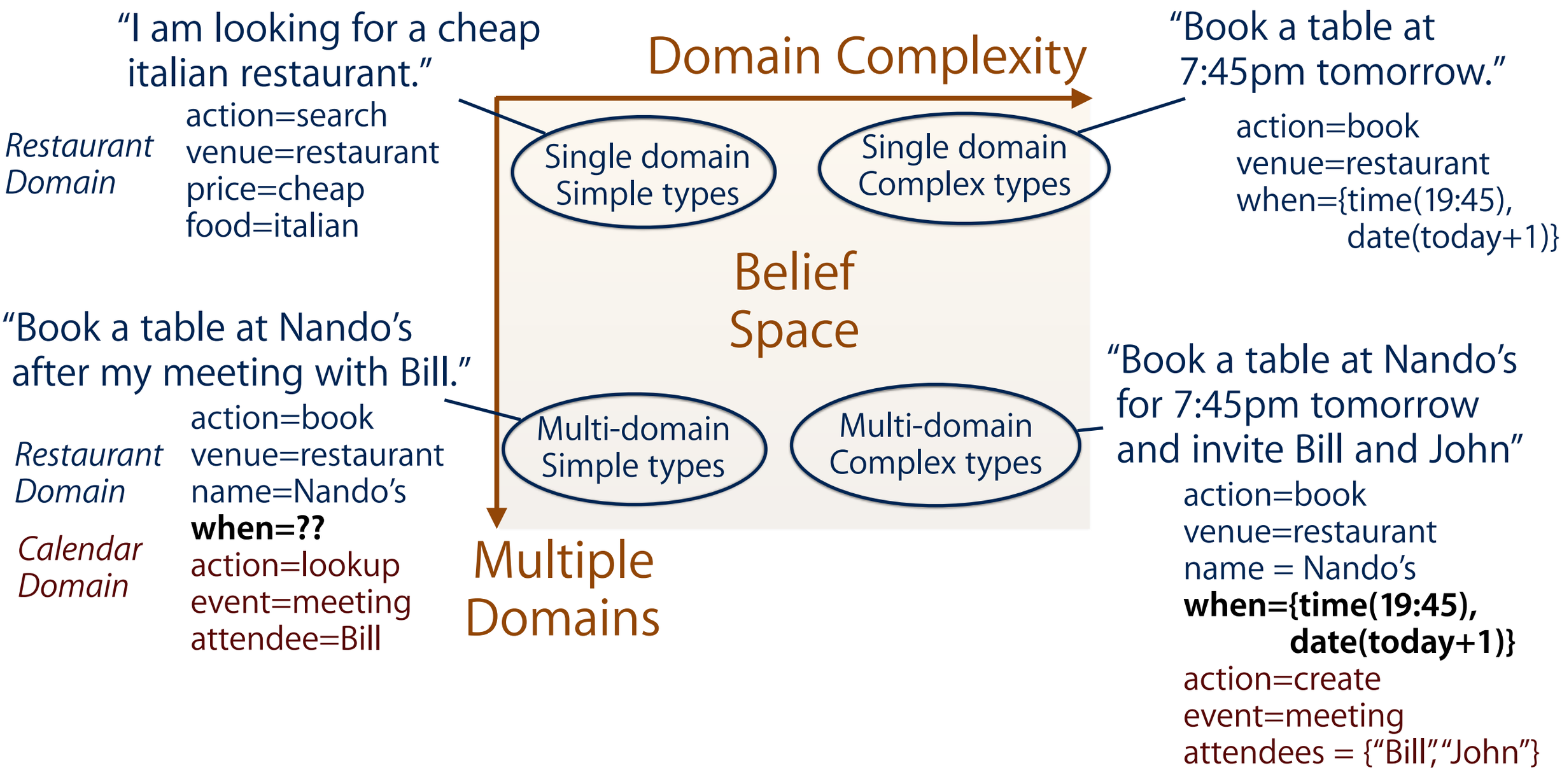
Real User Results

SL	94.5%
SL+RL	98.2%

NN Policy trained and tested on-line with real users.

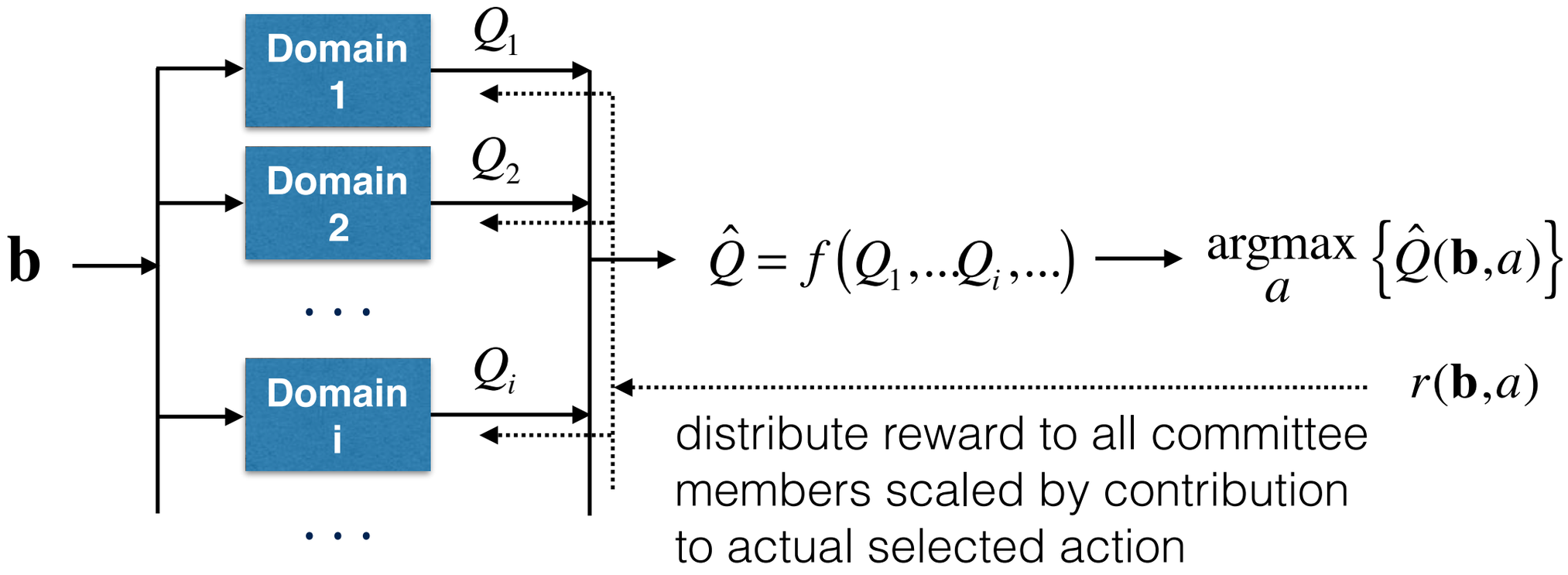
Su, P-H, et al., Continuously Learning Neural Dialogue Management, arXiv:1606.02689

Curse of Dimensionality



Bayesian Committee Machines

Assume M independent policies and a common belief state



Example using GP-RL:

$$\hat{Q} \sim N\left(\bar{Q}, \Sigma^{Q}\right)$$

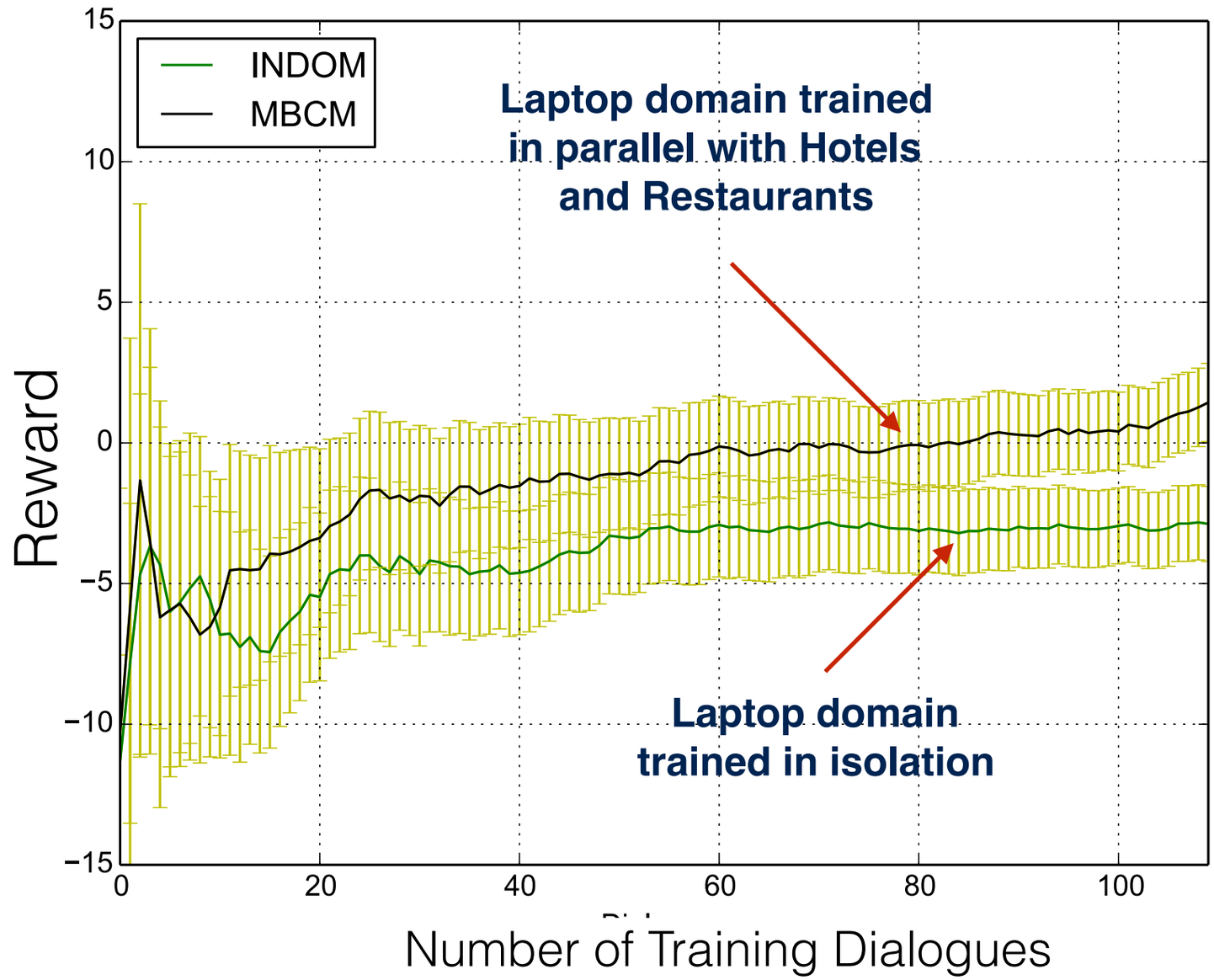
where

$$\bar{Q} = \Sigma^{Q} \sum_{i=1}^{M} \left(\Sigma_i^{Q}\right)^{-1} Q_i$$

$$\Sigma^{Q} = \left[\sum_{i=1}^{M} \left(\Sigma_i^{Q}\right)^{-1} - const\right]^{-1}$$

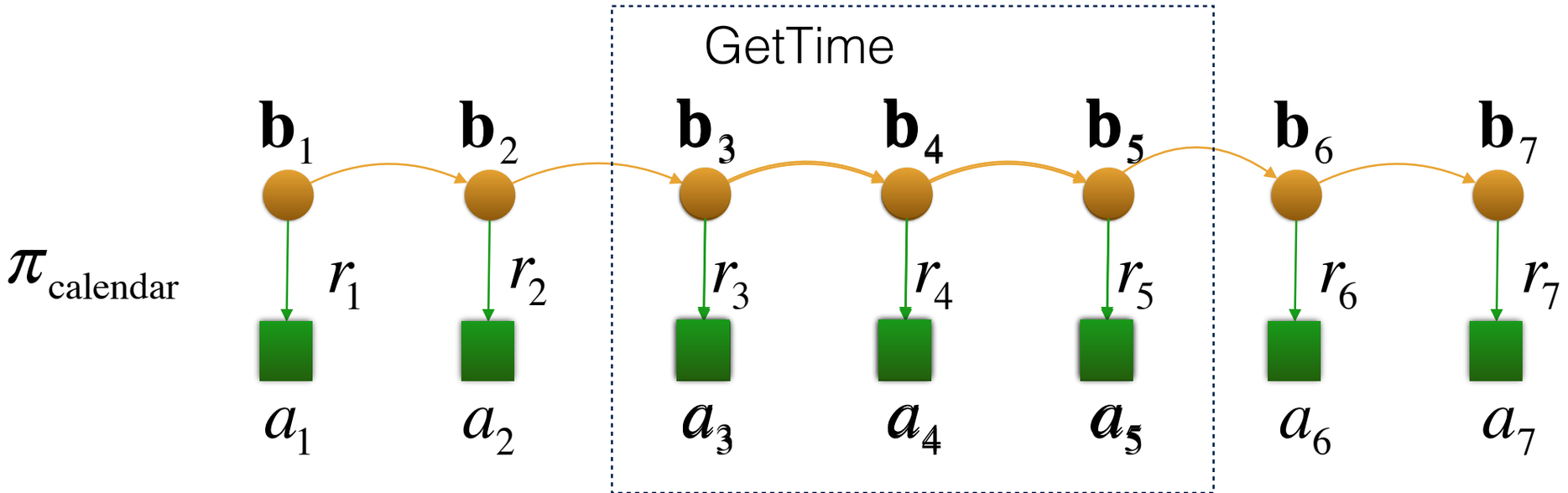
Three domains trained from scratch on line both individually and in parallel:

- Hotel info
- Restaurant info
- Laptop product guide



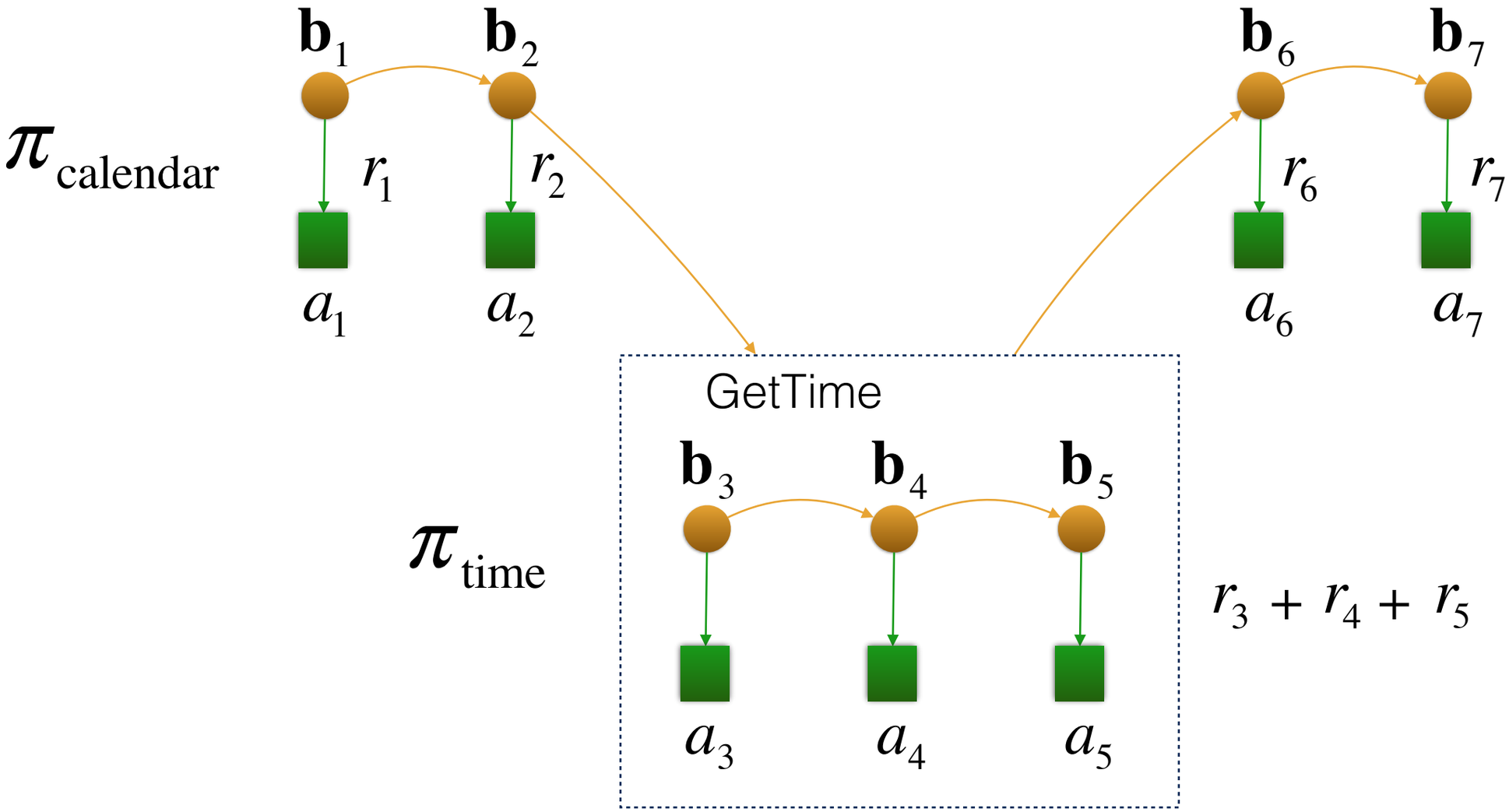
M. Gasic et al (2015). "Policy Committee for Adaptation in Multi-domain Spoken Dialogue Systems." IEEE ASRU 2015, Scotsdale, AZ.

Domain Complexity

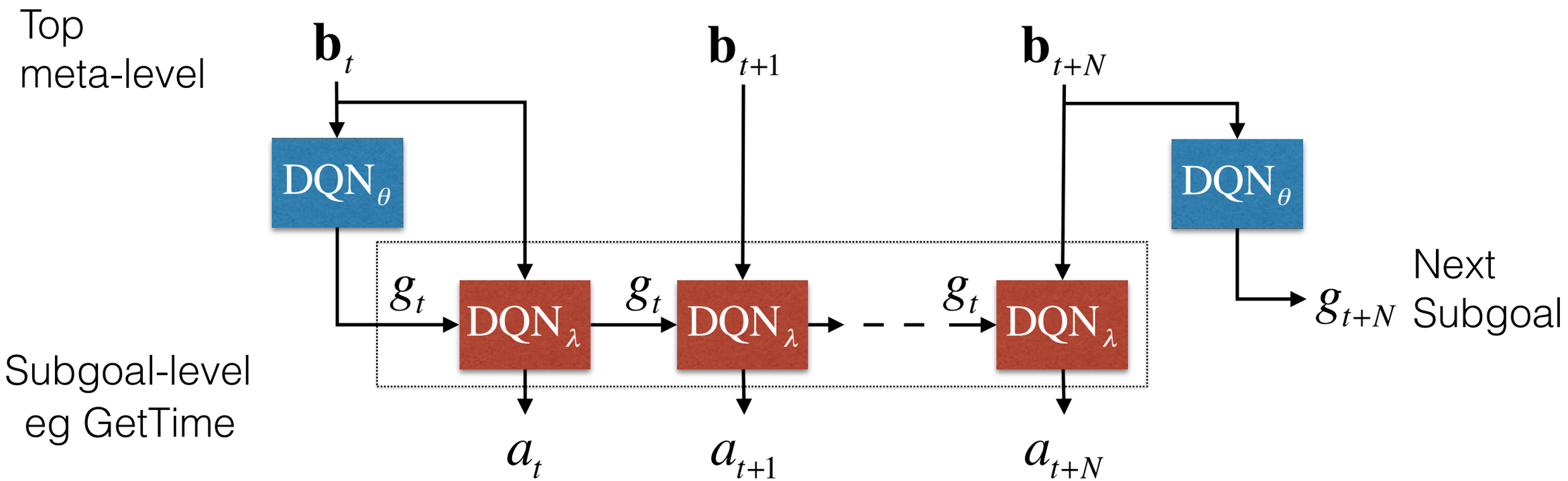


System:	How can I help?	Who with?	What time?	Was that 9.30?	5.30 pm?	Ok meeting at 5.30pm with Bill?	Meeting is scheduled
User:	Fix a meeting	Bill	5.30	No, 5.30	Yes	Yes	

Hierarchical Reinforcement Learning



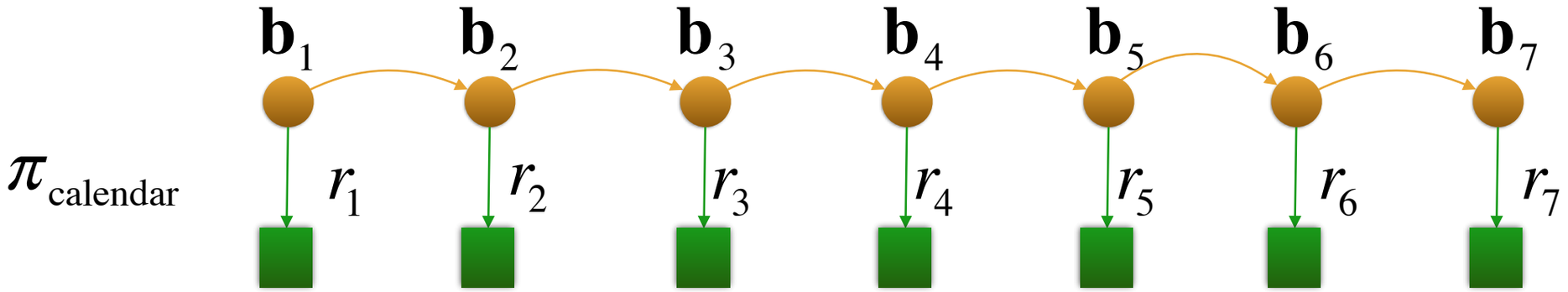
Hierarchical Deep Reinforcement Learning



T. Kulkarni et al (2016). "Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation." arXiv:1604.06057.

Measuring Success

Task success is not always obvious….

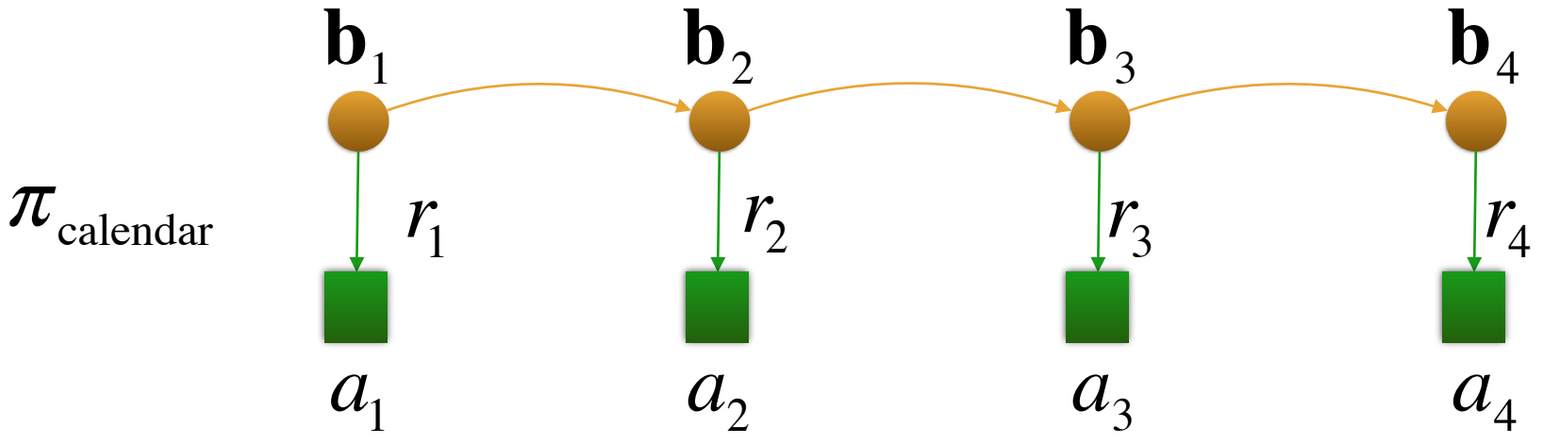


	a_1	a_2	a_3	a_4	a_5	a_6	a_7
System:	How can I help?	Who with?	What time?	Was that 9.30?	5.30 pm?	Ok meeting at 5.30pm with Bill?	Meeting is scheduled
User:	Fix a meeting	Bill	5.30	No, 5.30	Yes	Yes	✔

….so probably ok

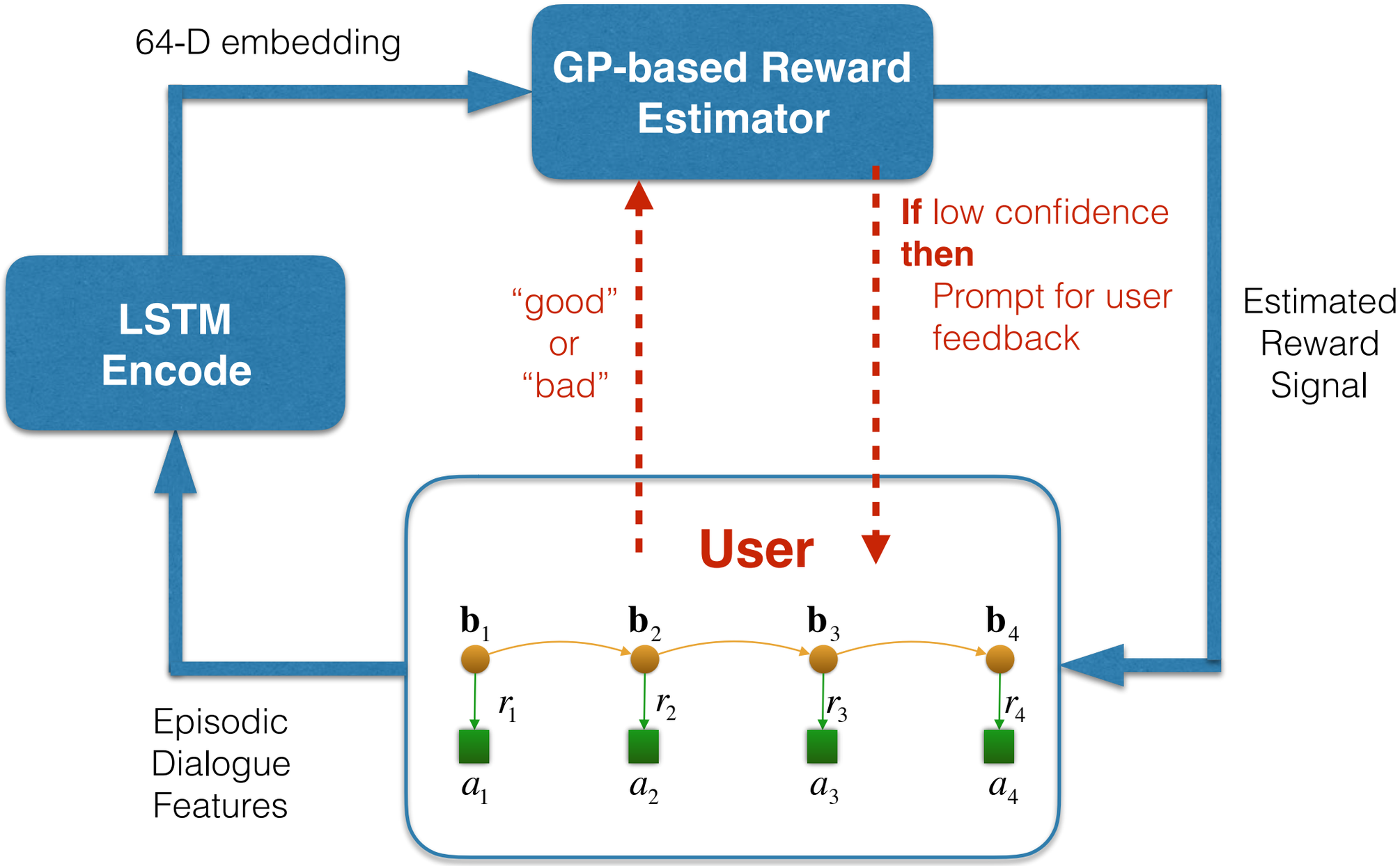
Measuring Success

However, what about the problematic weather query?

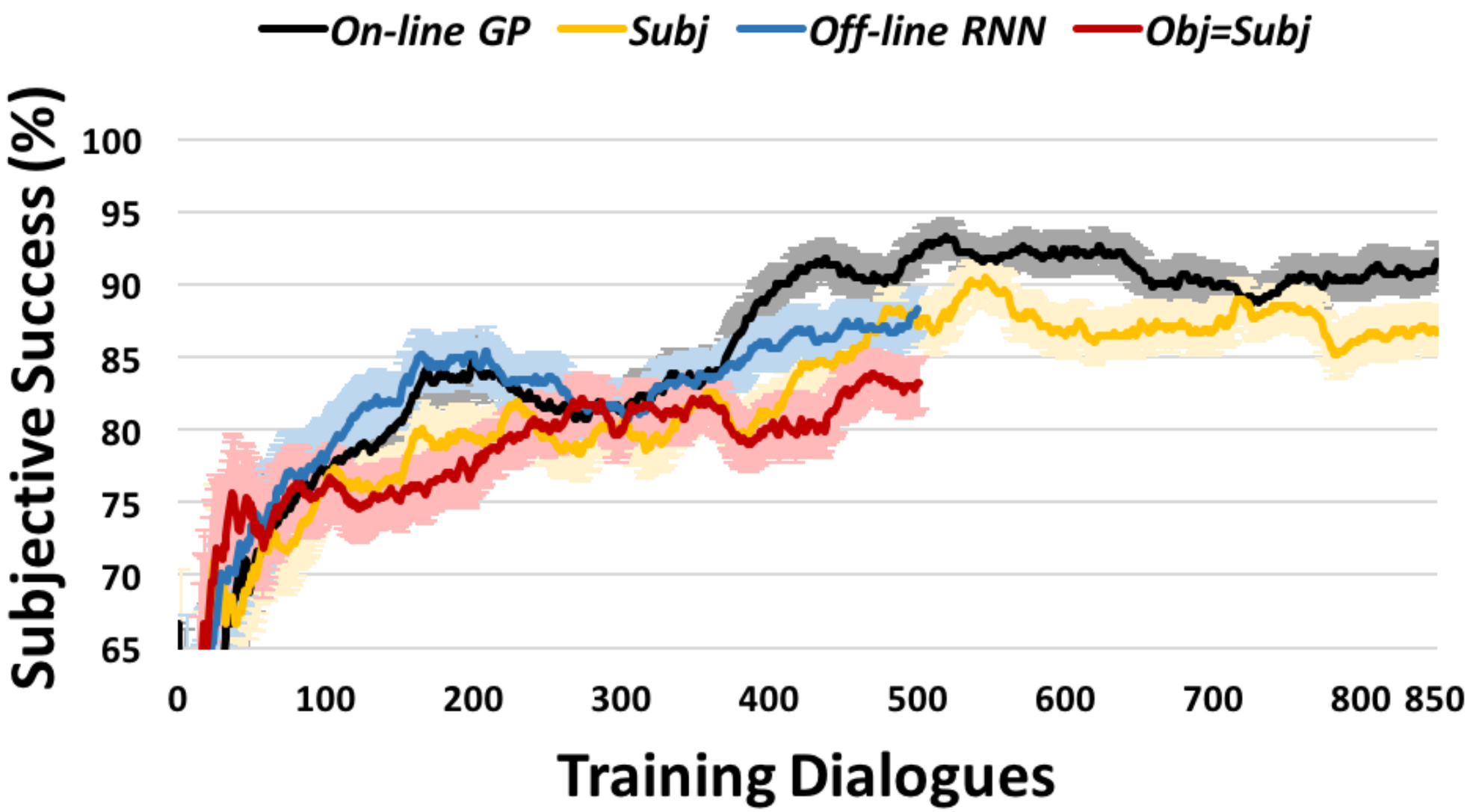


System:	How can I help?	It'll be fine all day in the Bay area.	I know your name Steve, it's "Steve".	I dont believe it's raining right now.
User:	Hows the weather in Maine	No, Maine	I want the weather in Maine!	

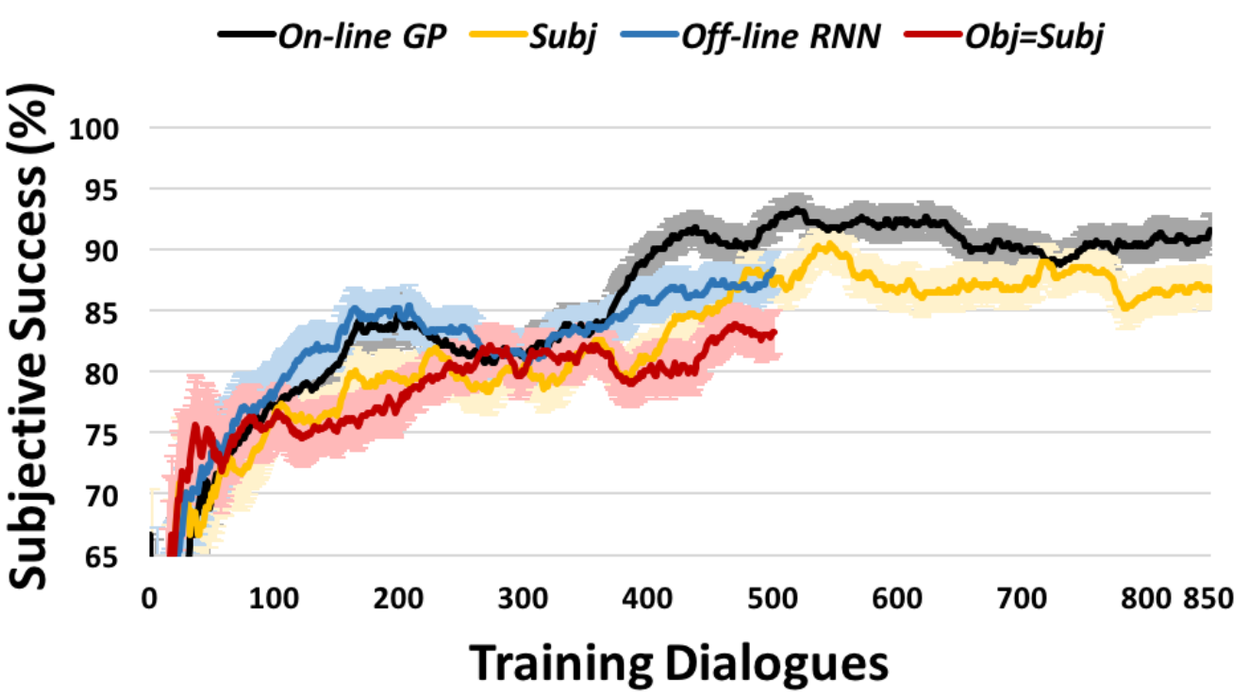
On-line Reward Estimation

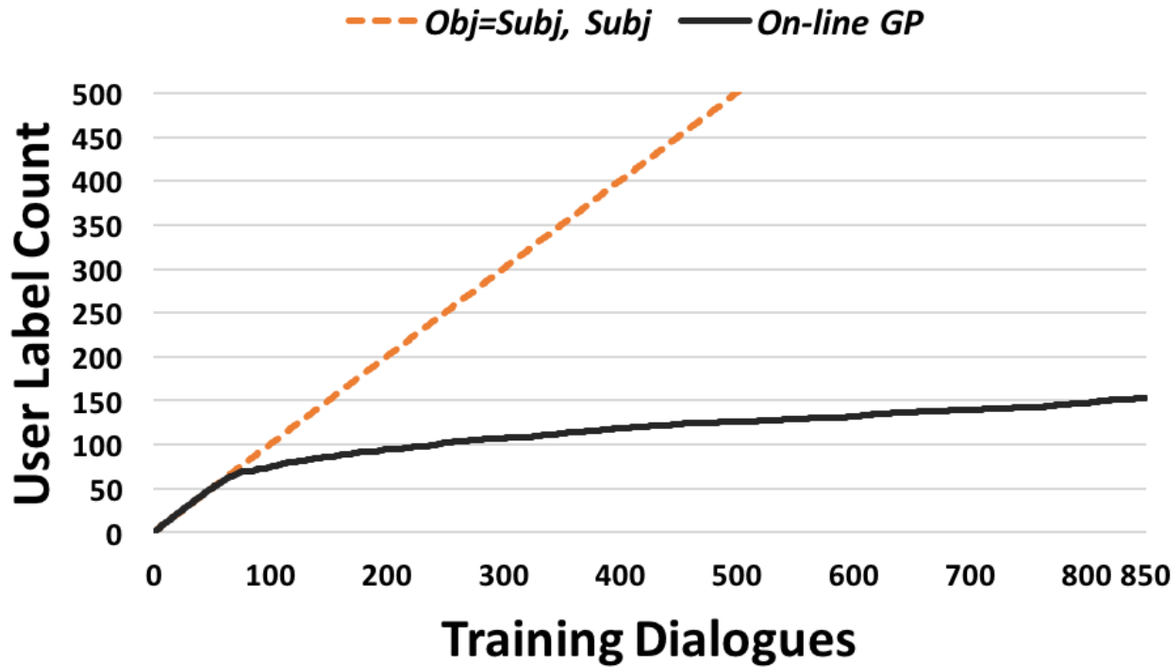


On-line Reward and Policy Learning



On-line Reward and Policy Learning





P-H. Su et al (2016). "On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems." ACL 2016, Berlin.

Summary

- POMDPs and Reinforcement Learning provide a powerful mathematical framework for decision making in intelligent conversational agents.
- DNNs provide a flexible building block for all stages of the dialogue system pipeline, though training is often problematic.
- Unrestricted conversation is challenging but there are several promising approaches to managing complexity.
- For commercially deployed systems, the user is a tremendous untapped resource, and Reinforcement Learning provides the framework for exploiting it.

Credits

All members of the Cambridge Dialogue Systems Group Past and Present:

Milica Gasic
Catherine Breslin
Pawel Budzianowski
Matt Henderson
Filip Jurcicek
Simon Keizer
Dongho Kim
Fabrice Lefevre
Francois Mairesse
Nikola Mrksic
Lina Rojas Barahona
Jost Schatzmann
Matt Stuttle
Martin Szummer
Eddy Su
Blaise Thomson
Pirros Tsiakoulis
Stefan Ultes
David Vandyke
Karl Weilhammer
Shawn Wen
Jason Williams
Hui Ye
Kai Yu