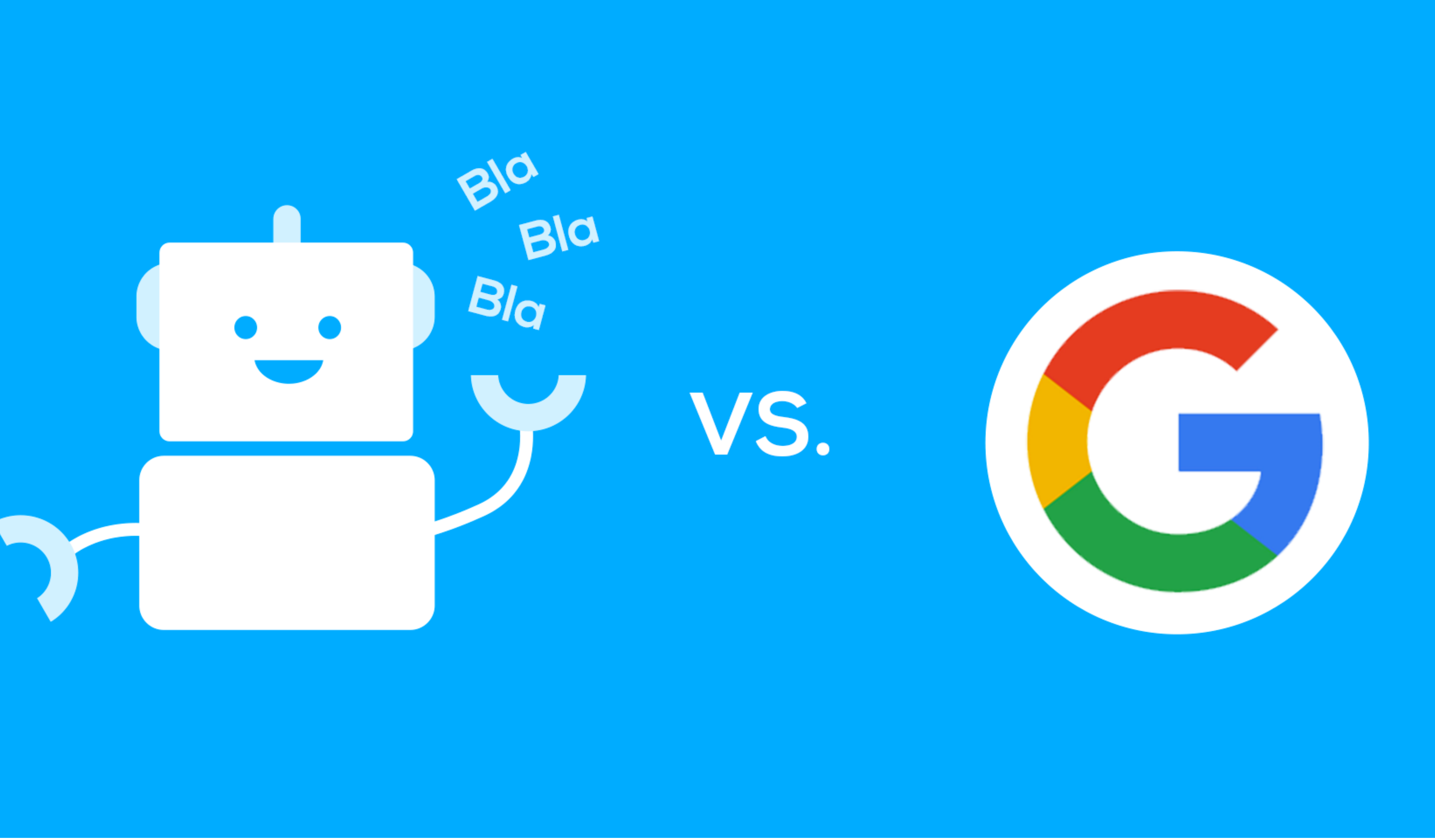
# Reinforcement Learning Approaches in Dialogue System and Chatbots

## Head First Theory and Practice

Yanran Li
The Hong Kong Polytechnic University



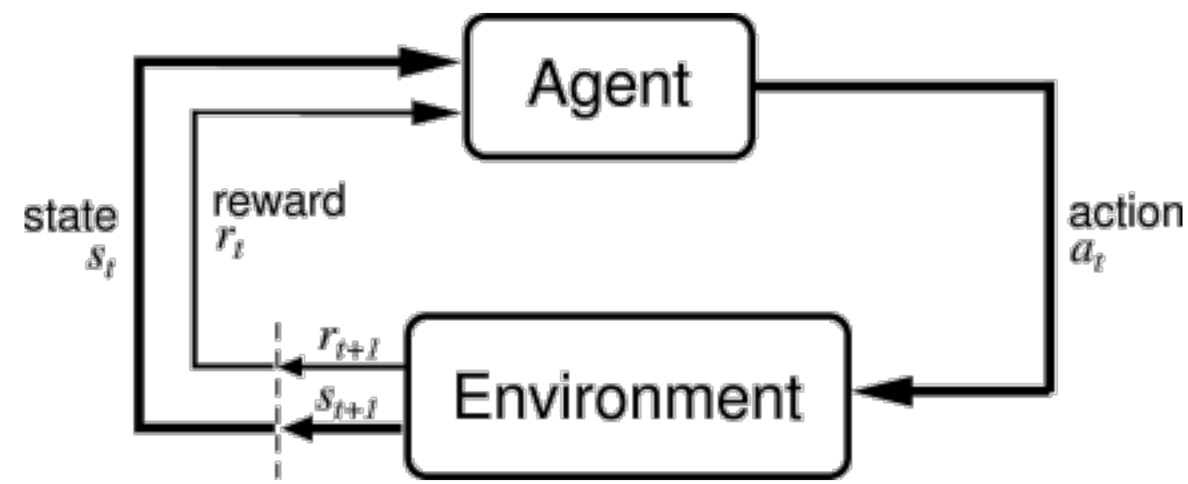

# Introduction | to Reinforcement Learning

# Reinforcement Learning

- Learn to make good sequences of decisions

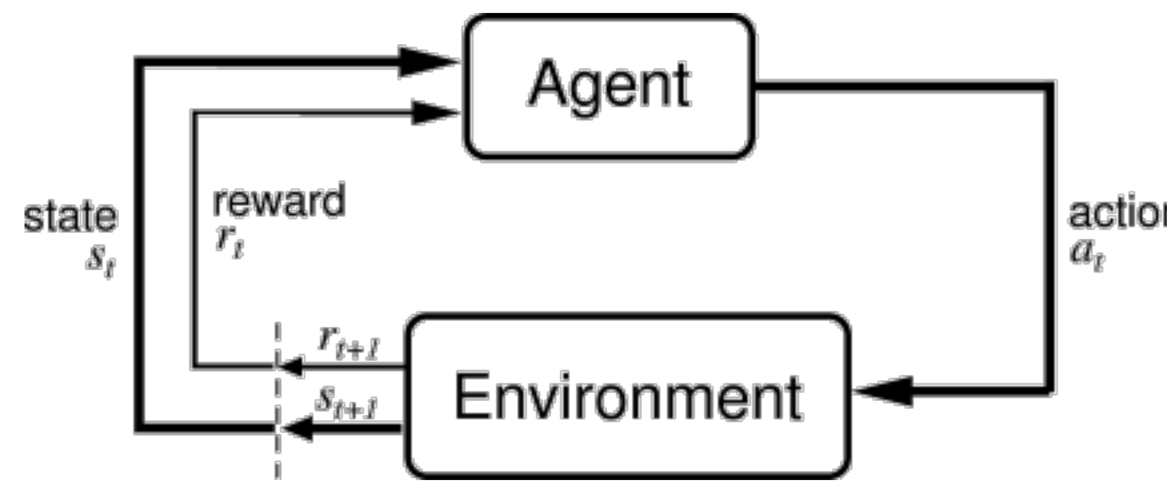
# Reinforcement Learning

- Learn to make good sequences of decisions



# Reinforcement Learning

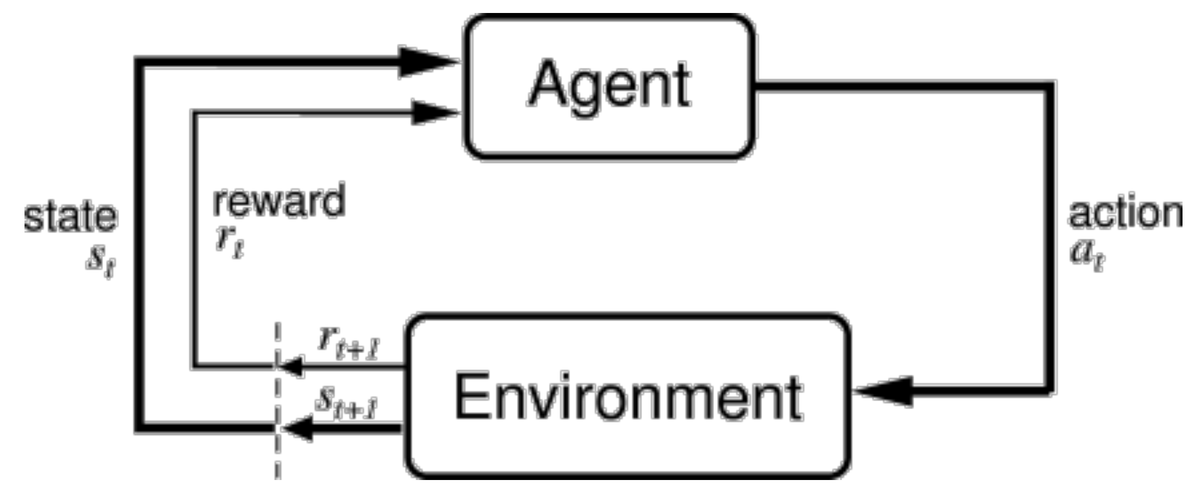
- Learn to make good sequences of decisions
- Policy: mapping from history of past actions, states, rewards to next action




- S: set of states
- A: set of actions
- R: reward model R(s) / R(s,a) / R(s,a,s')
- T: dynamics model
- ϒ: discount factor

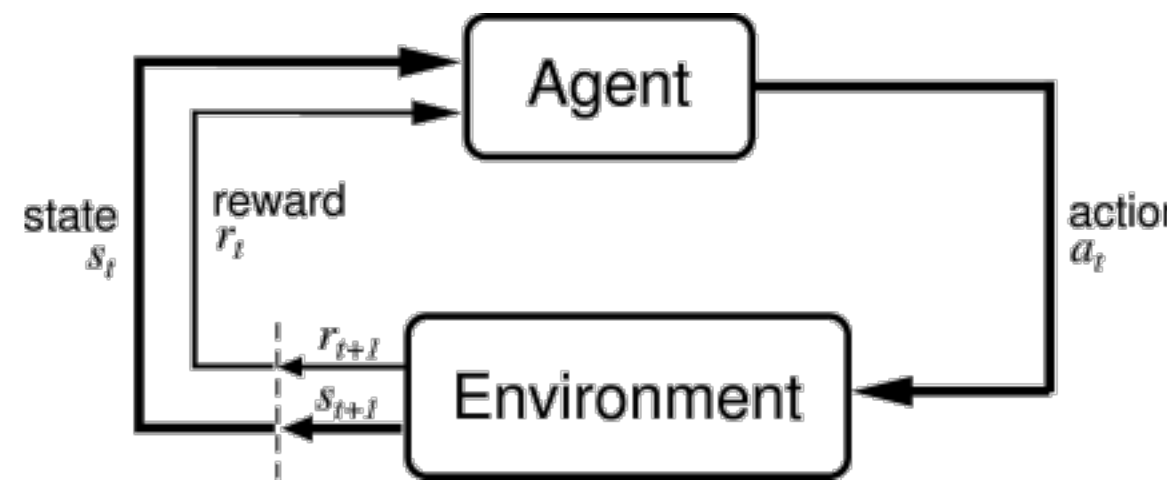
# Reinforcement Learning

- At each step t the agent:
  - Executes action
  - Receives observation
  - Receives scalar reward
- The environment:
  - Receives action at
  - Emits observation
  - Emits scalar reward



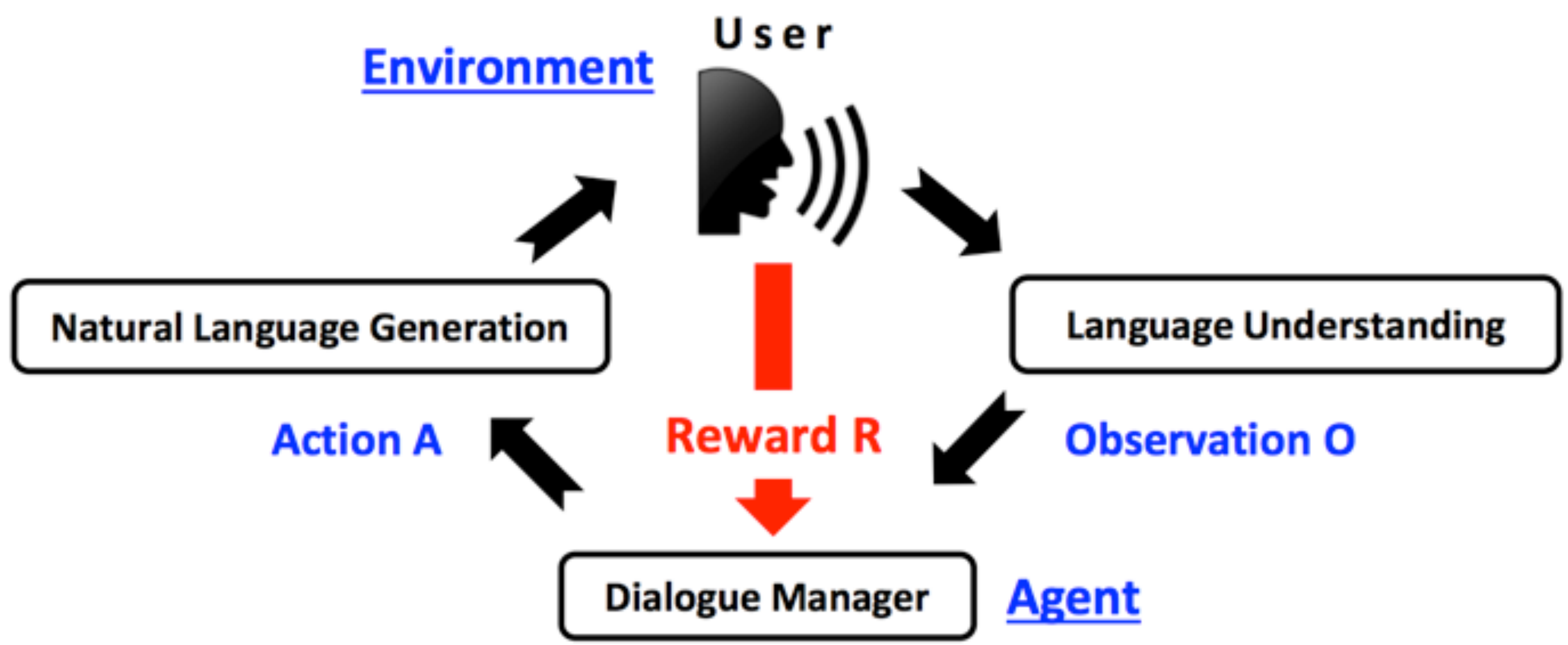
# Reinforcement Learning

- At each step t the agent:
  - Executes action
  - Receives observation
  - Receives scalar reward
- The environment:
  - Receives action at
  - Emits observation
  - Emits scalar reward



# Reinforcement Learning in Dialogue Setting

- Optimized dialogue policy selects the best action that can maximize the future reward. Correct rewards are a crucial factor in dialogue policy training.



# Reinforcement Learning in Dialogue Setting

- Observation / action
  - Raw utterance (natural language form)
  - Semantic representation (dialog-acts)
- Reward
  - +10 upon termination if succeeded
  - −10 upon termination if failed o
  - −1 per turn
- State
  - Explicitly defined (POMDP-based, …)
  - Implicitly defined (RNNs)

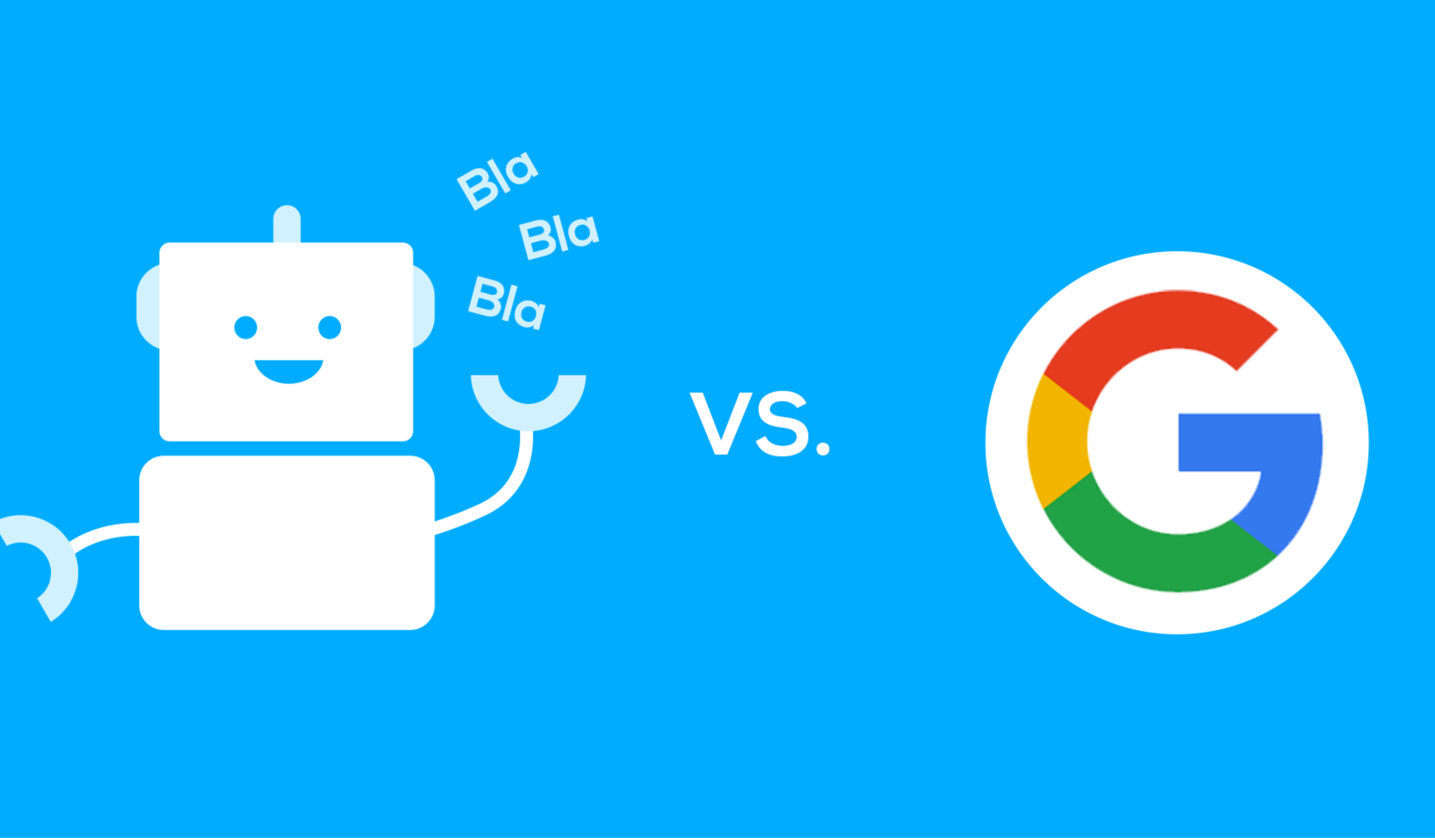
# Supervised Learning V.S. Reinforcement Learning

- Distinct methods of learning from experience
- SL – learning from **previous experience**
    - Learning a model on collected input-output pairs (training data)
    - by minimizing some loss functions
    - No explicit dependence on how training data is collected
- RL – learning by **experiencing**
    - An agent learned by interacting with an environment to achieve a goal
    - Learning by trial and error (exploration) with only delayed reward
    - Can tell for itself when it is right or wrong
    - RL is more realistic, natural and ambitious than SL

# Reinforcement Learning Agent

| Input message | Supervised Learning Agent | Reinforcement Learning Agent |
|---|---|---|
| How old are you ? | I'm 16. | I'm 16, why are you asking ? |
| What's your full name ? | I have no idea . | What's yours ? |
| I don't want to go home tonight. | Really ? | Why ? |
| Do you have any feelings for me ? | I don't know what you are talking about. | I don 't want to hurt your feelings . |
| How much time do you have here? | Not long enough. Sorry, sir. | Ten seconds . |
| Shall we get started ? | Of course ! | Yes . We 've got a lot of work to do here . |
| Do you play football ? | No, i don't | Yes. I love football ! |
| We'd have to talk to him. | I mean, he's a good guy | About what ? |
| How come you never say it? | Because I don't want to hurt you . | I don 't think it 's a good idea to say it . |

- RL agent generates more interactive responses
- RL agent tends to end a sentence with a question and hand the conversation over to the user




KB-InfoBot

# Three Types of Dialogue Systems

Specific Knowledge

↑

General Knowledge

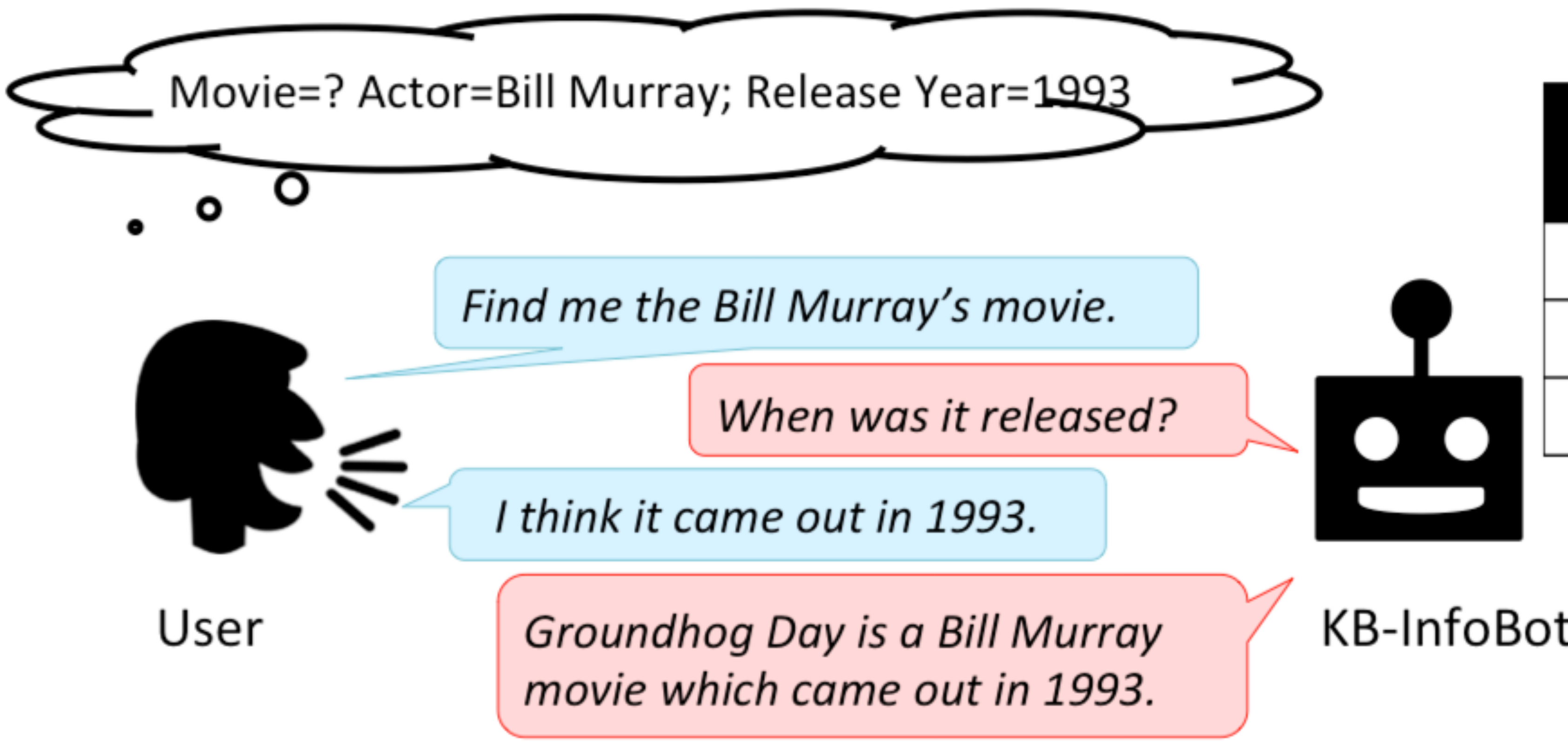
| | |
|---|---|
| Task-Oriented Dialogue System (TASK COMPLETION) | • Slot Filling<br>• Dialogue Management<br>• Multi-turn (often) |
| InfoBot (INFORMATION & ANSWERS) | • Intelligent QA<br>• Intelligent Search<br>• Single-turn (most of time) |
| ChatBot (GENERAL CONVERSATION) | • Chit-chat<br>• Emotion/Sentiment Analysis<br>• Multi-turn (often) |

[slides from Nan Duan]

# Three Types of Dialogue Systems

- Task-completion bot
  - Movie ticket booking
  - Hotels booking
  - Travel assistant
- Info bot
  - Find the closest Starbucks with drive-thru
  - Find a family-friendly movie directed by Andrew Stanton near Redmond for upcoming weekend afternoons

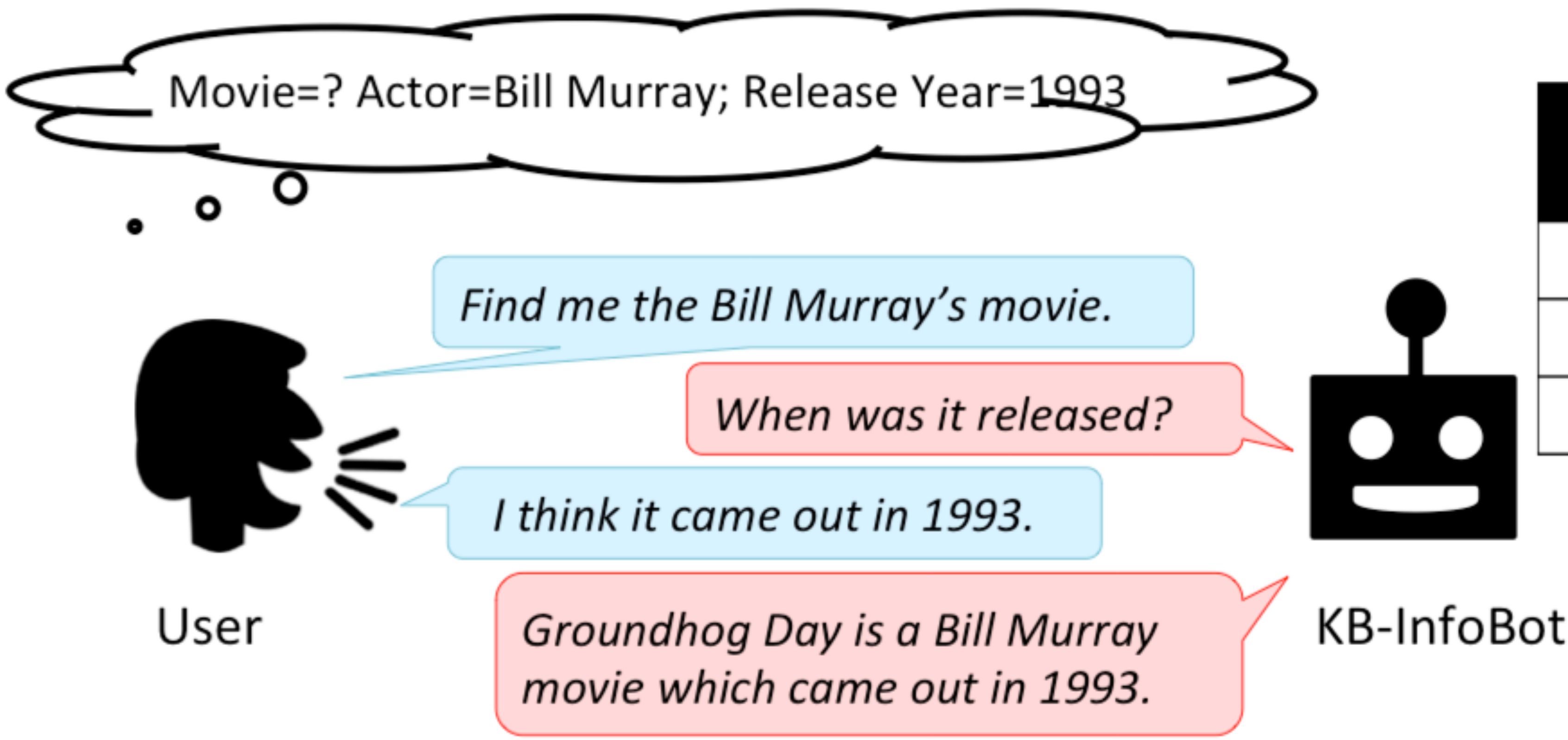
# KB-InfoBot




**Entity-Centric Knowledge Base**

| Movie | Actor | Release Year |
|---|---|---|
| *Groundhog Day* | Bill Murray | 1993 |
| *Australia* | Nicole Kidman | X |
| *Mad Max: Fury Road* | X | 2015 |

# KB-InfoBot

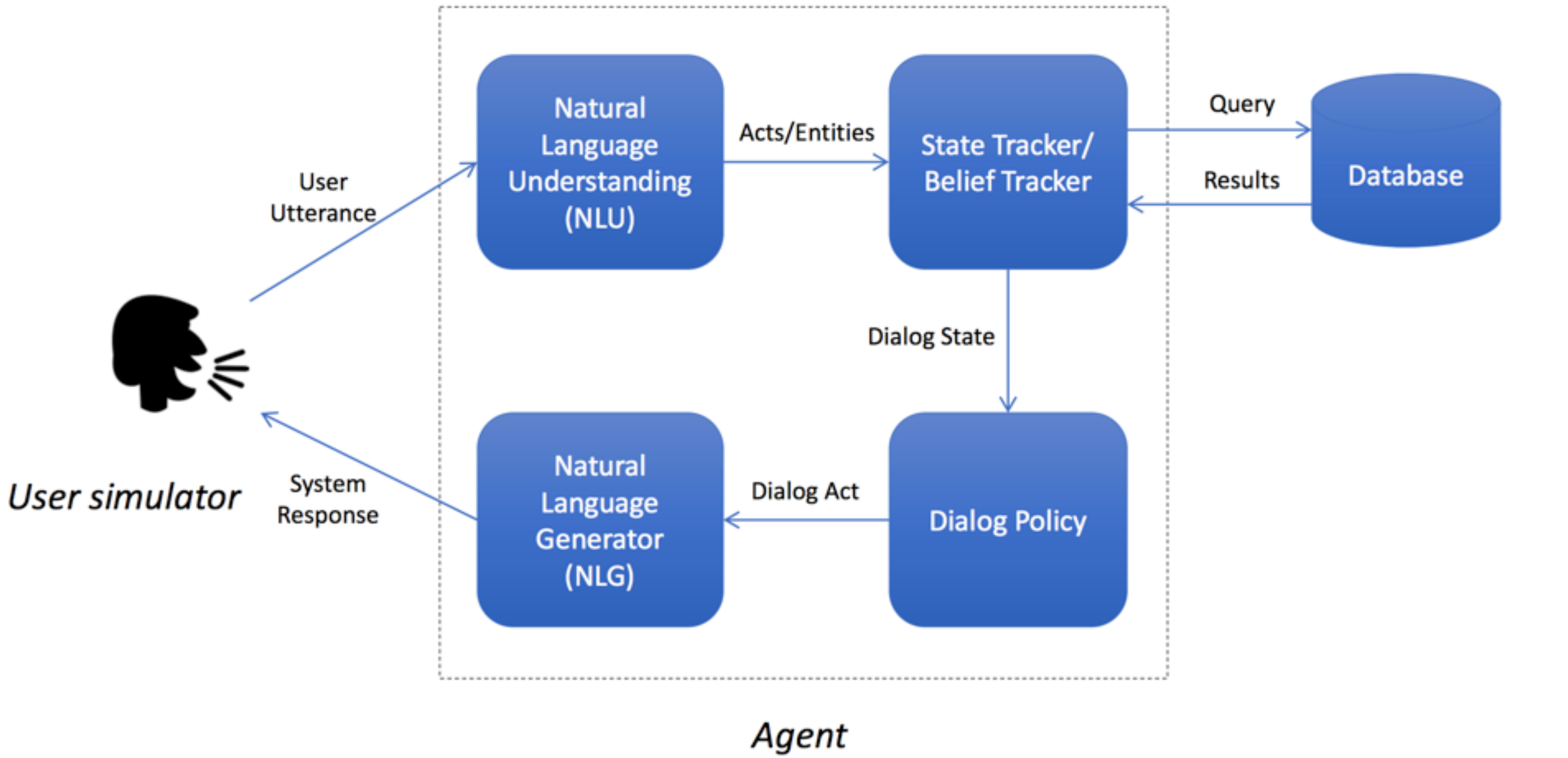



**Entity-Centric Knowledge Base**

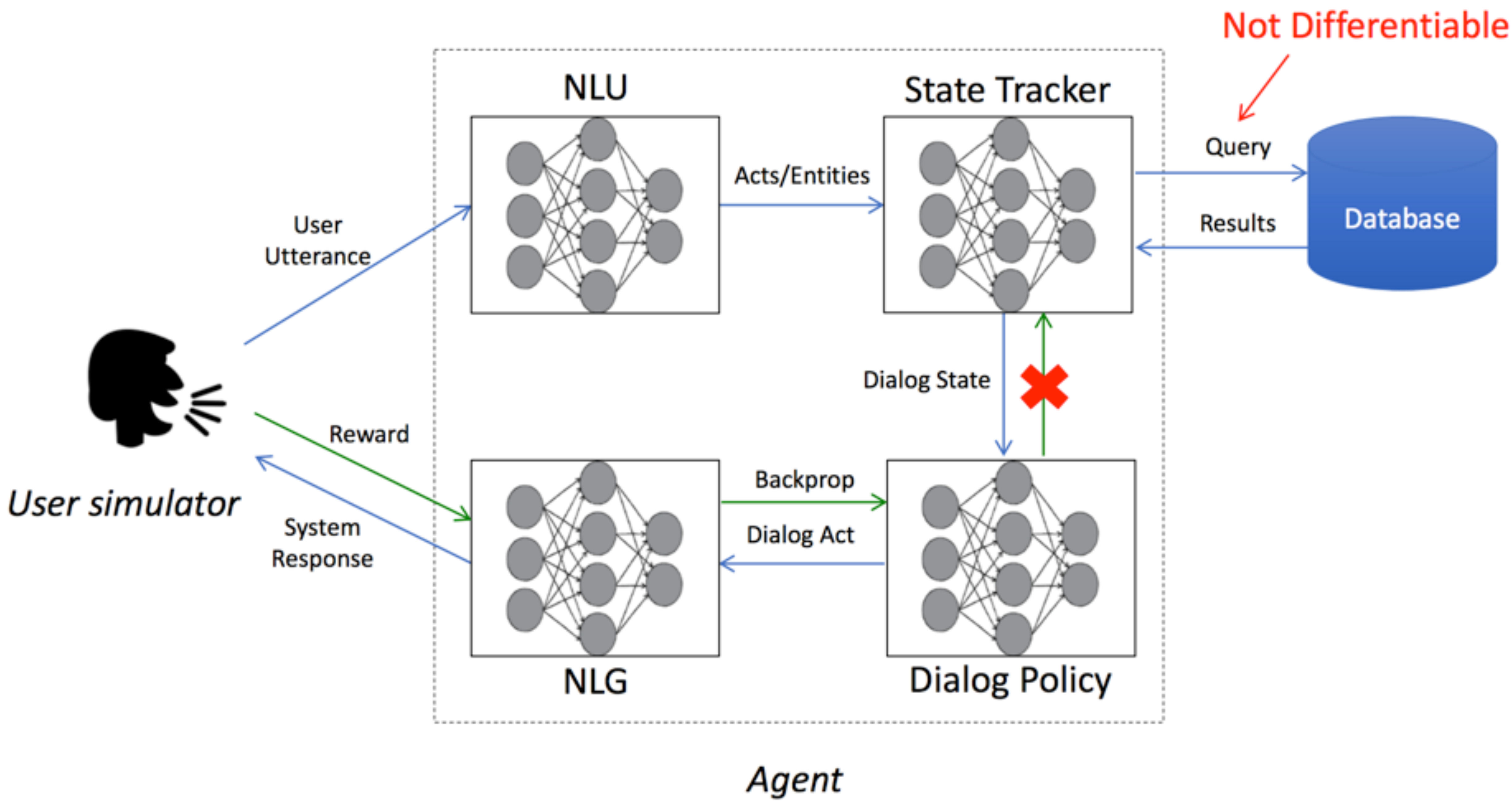
| Movie | Actor | Release Year |
|---|---|---|
| *Groundhog Day* | Bill Murray | 1993 |
| *Australia* | Nicole Kidman | X |
| *Mad Max: Fury Road* | X | 2015 |

- Setting:
    - User is looking for a piece of information from one or more tables/KBs
    - System must iteratively ask for user constraints ("slots") to retrieve the answer

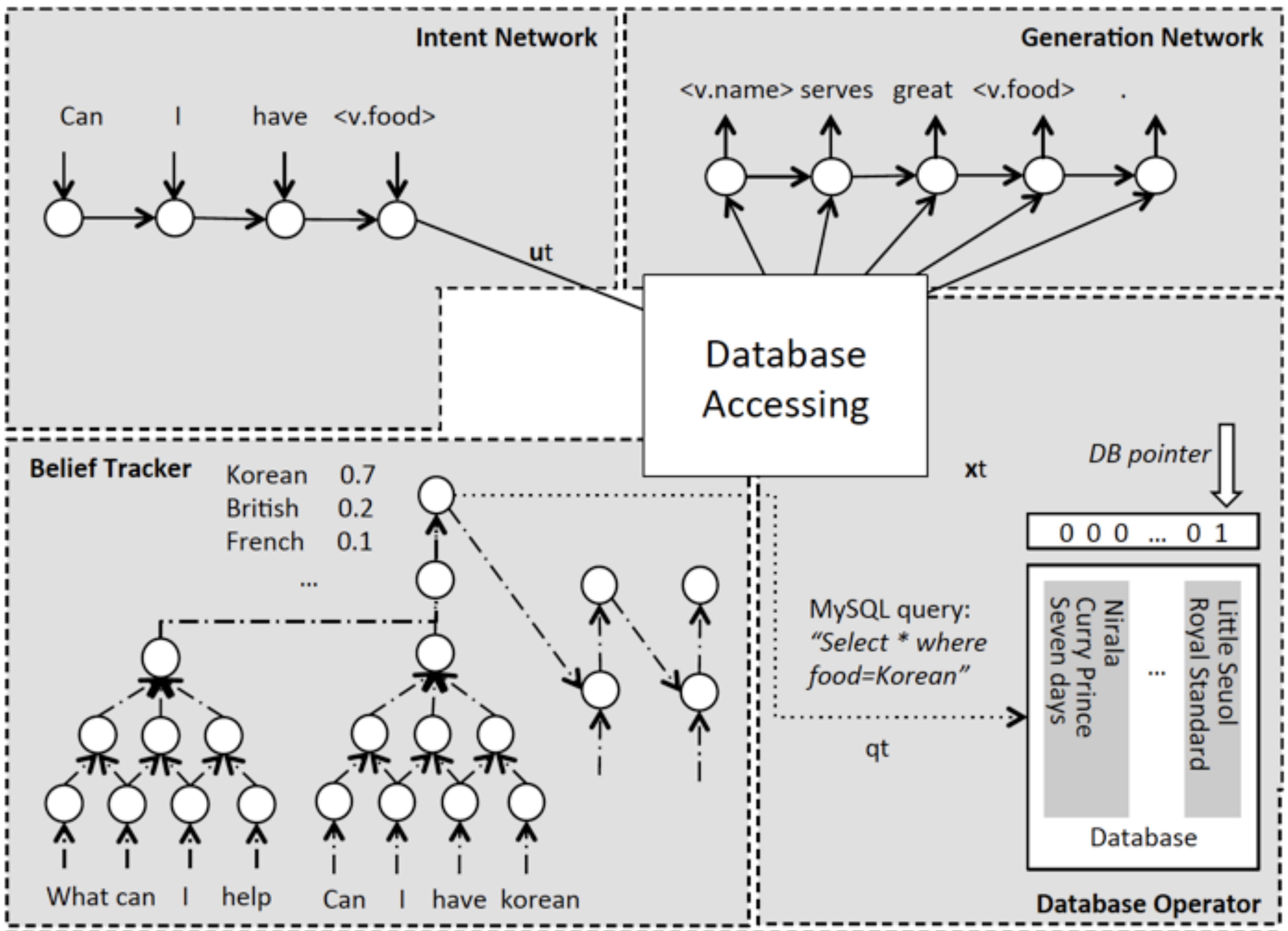
# Task-oriented Dialogue System



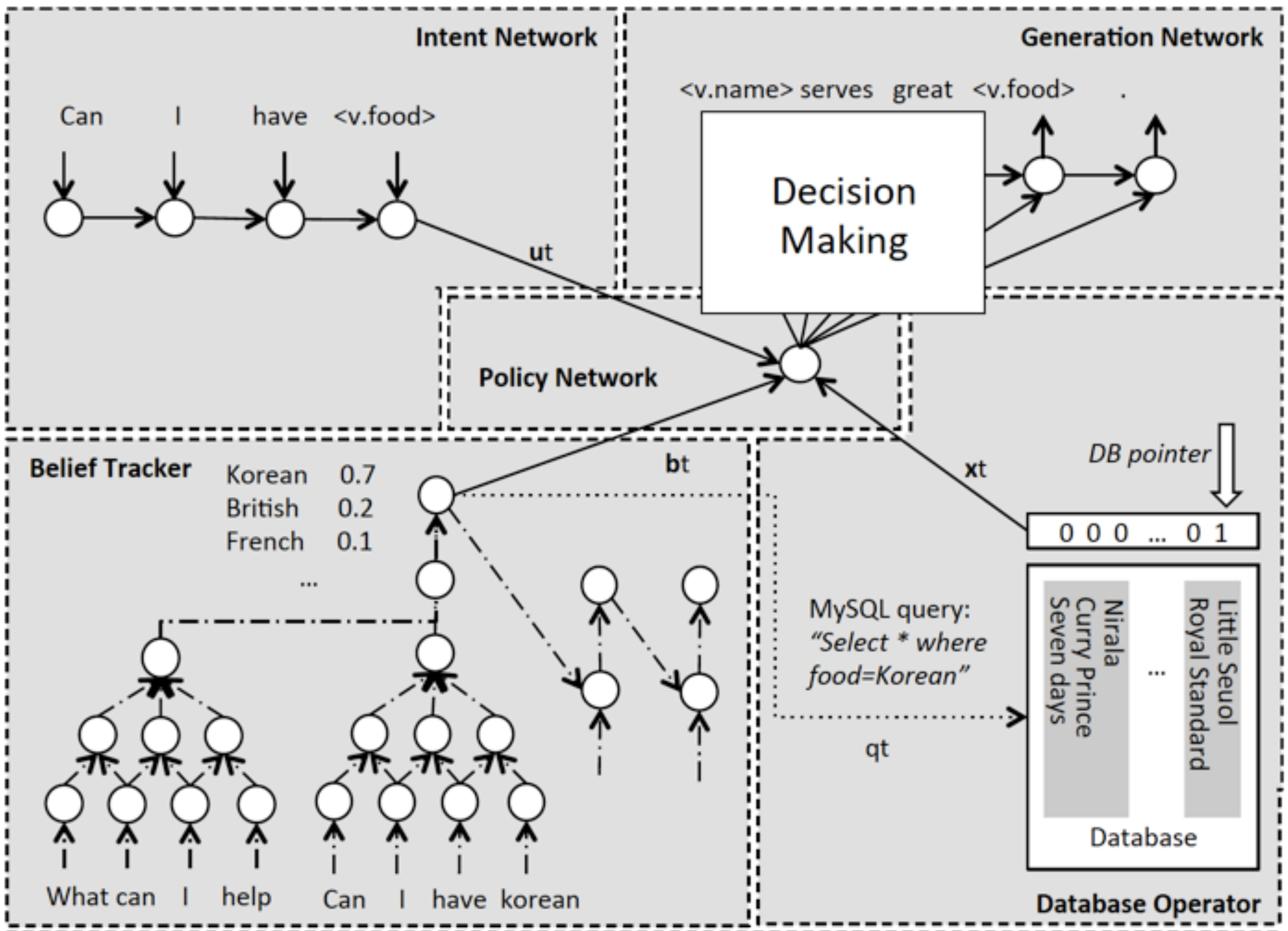
# Task-oriented Dialogue System



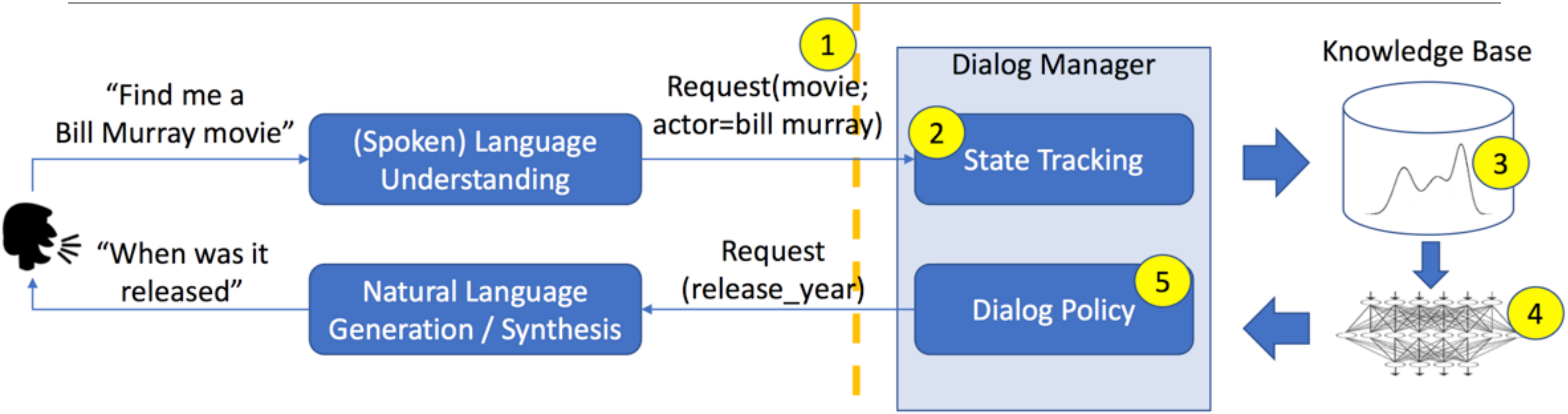
# Task-oriented Neural Dialog System



# Task-oriented Neural Dialog System



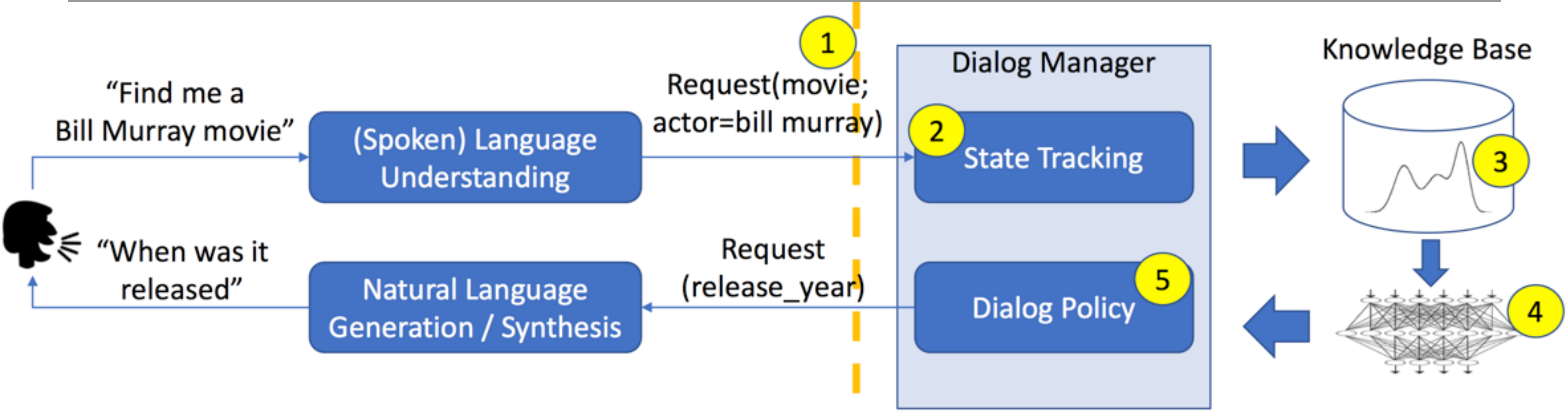
# Task-oriented Dialogue System




1. Use a single deep NN for {dialog manager and KB}

2. Recurrent network to track states of conversation

# Task-oriented Dialogue System

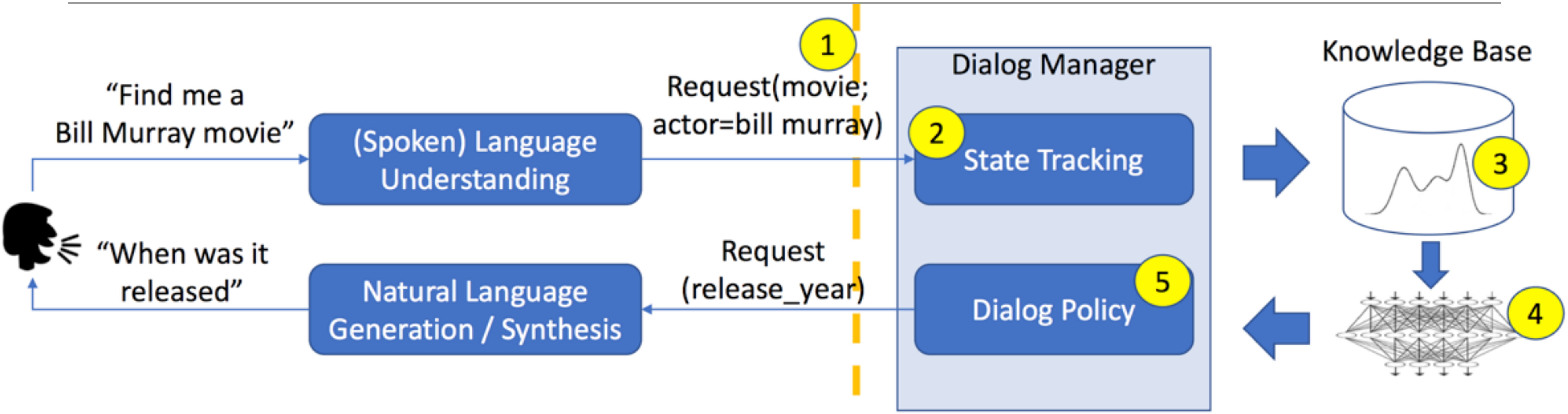



1. Use a single deep NN for {dialog manager and KB}

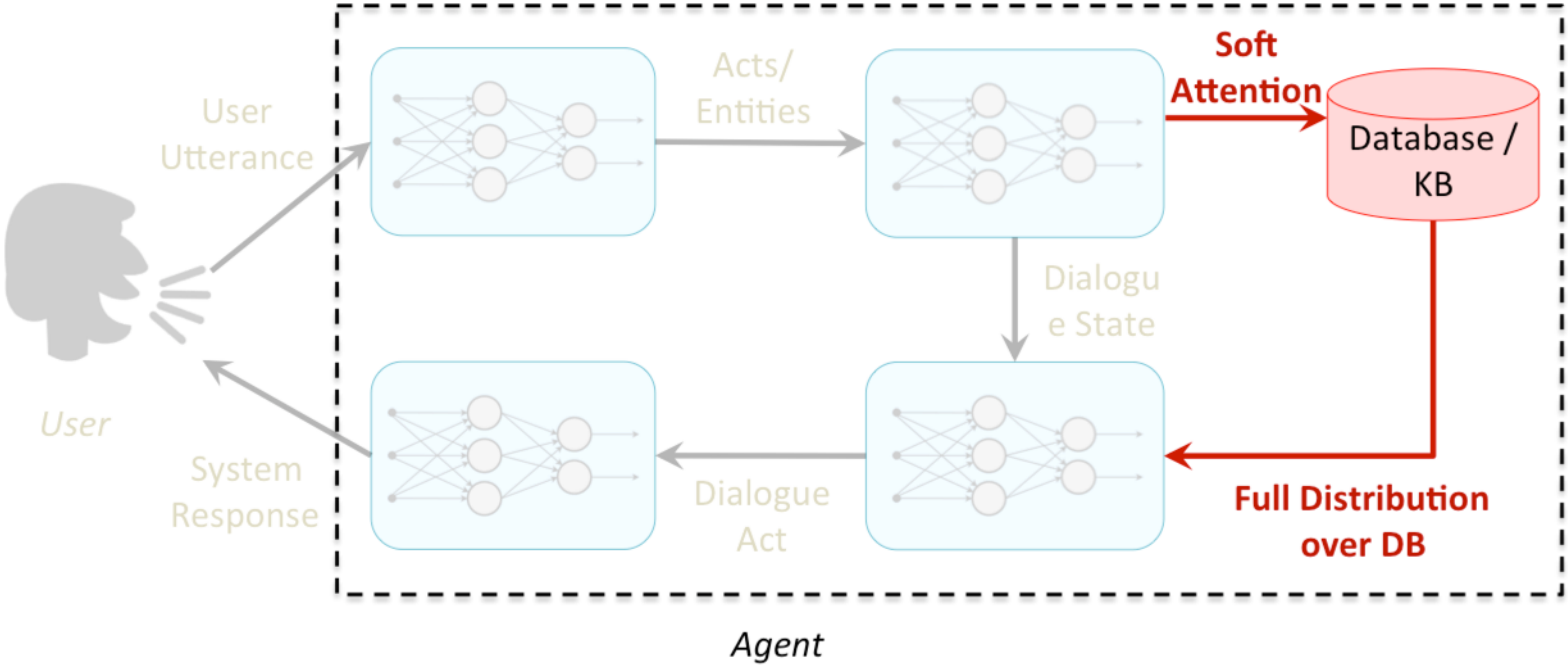
2. Recurrent network to track states of conversation

3. **Maintain (implicitly) a distribution over entities in KB**

# Soft-KB Lookup via Attention

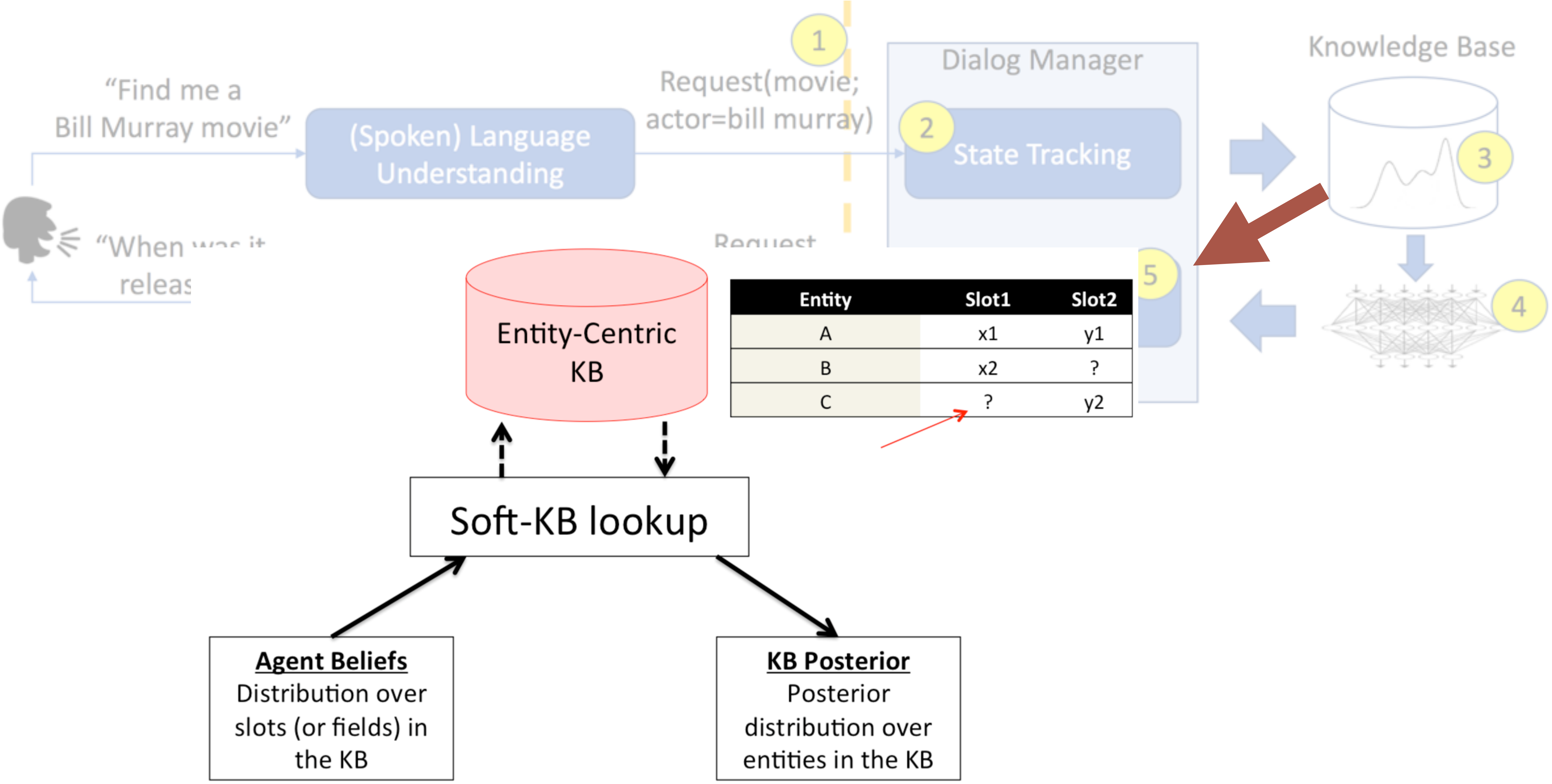



- Replace symbolic query with an attention distribution
  - Compose slot-wise belief states into one posterior distribution over entire database
  - The KB structure is encoded in the computation of attention



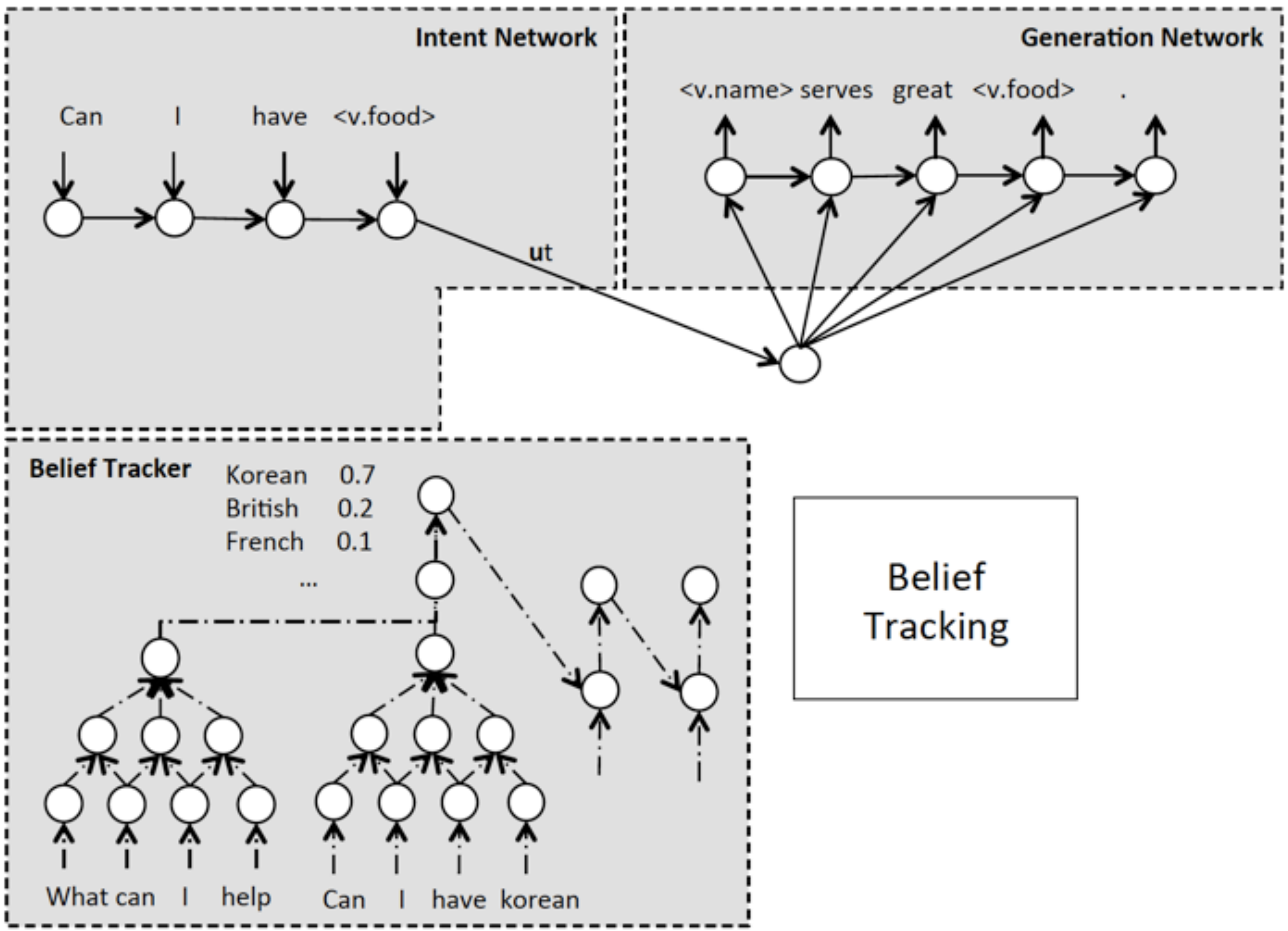

- Replace symbolic query with an attention distribution
  - Compose slot-wise belief states into one posterior distribution over entire database
  - The KB structure is encoded in the computation of attention

# Soft-KB Lookup via Attention




| Entity | Slot1 | Slot2 |
|---|---|---|
| A | x1 | y1 |
| B | x2 | ? |
| C | ? | y2 |

# State/Belief Tracker




Belief Tracking

# Agent Beliefs via State Tracker

- For each slot j:
  - A multinomial over slot values –

$$p_j^t(v)$$

| Slot Values → | x1 | x2 |
|---|---|---|
| Probabilities → | 0.3 | 0.7 |

  - A binomial probability of whether user knows the value of the slot -

$$q_j^t \quad \boxed{0.8}$$

# KB Posterior

| Entity | Slot1 | Slot2 |
|---|---|---|
| A | x1 | y1 |
| B | x2 | ? |
| C | ? | y2 |

$$\Pr(\text{Entity}) \propto \Pr(\text{Entity-Slot1}) \times \Pr(\text{Entity-Slot2})$$

# KB Posterior

$$
\begin{aligned}
&\Pr(\text{Entity-Slot1}) = \\
&\qquad \Pr(\text{User knows Slot1}) \times \boxed{\Pr(\text{Entity-Slot1}|\text{User knows Slot1})} \\
&+ (1 - \Pr(\text{User knows Slot1})) \times \Pr(\text{Entity-Slot1}|\text{User does not know Slot1})
\end{aligned}
$$

$$\left[\; q_j^t \quad \boxed{0.8} \;\right] \qquad\qquad \left[\; \text{Uniform Prior} = \frac{1}{N} \;\right]$$

# KB Posterior

$$\Pr(\text{Entity-Slot1}|\text{User knows Slot1}) = \begin{cases} \Pr(\text{Known Values}) \times \dfrac{p_j^t(\text{Entity-Slot1-Value})}{\#\text{Entity-Slot1-Value}} \\ \Pr(\text{Missing Values}) \times \dfrac{1}{\#\text{Missing Values}} = \dfrac{1}{N} \end{cases}$$

Examples:

$$\Pr(\text{A-Slot1}|\text{User knows}) = \frac{2}{3} \times \frac{0.3}{1}$$

$$\Pr(\text{C-Slot1}|\text{User knows}) = \frac{1}{3}$$

| Entity | Slot1 |
|---|---|
| A | x1 |
| B | x2 |
| C | ? |

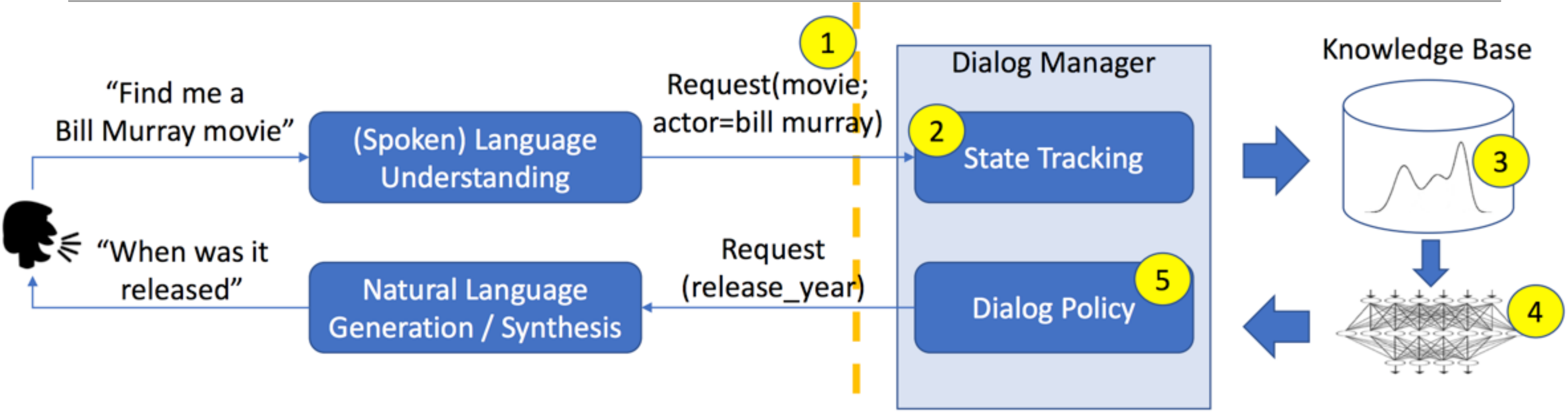
$p_j^t(v)$

| x1 | x2 |
|---|---|
| 0.3 | 0.7 |

# KB Posterior

- Distribution over all entities in the database
- Posterior reflects **uncertainty** in LU + State Tracking
- All operations are differentiable
  - Gradients can pass through during backward pass

# Task-oriented Dialogue System




1. Use a single deep NN for {dialog manager and KB}

2. Recurrent network to track states of conversation

3. **Maintain (implicitly) a distribution over entities in KB**

4. A summary network to "summarize" distribution information

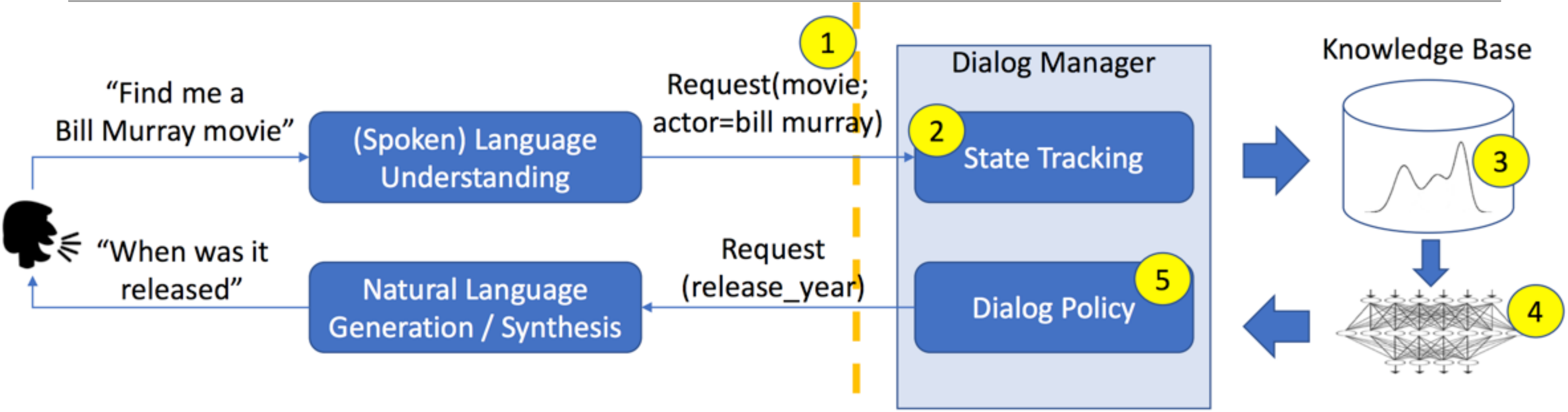
# Soft-KB Lookup

**Entity-Centric Knowledge Base**

| Movie | Actor | Release Year |
|---|---|---|
| Groundhog Day | Bill Murray | 1993 |
| Australia | Nicole Kidman | X |
| Mad Max: Fury Road | X | 2015 |

- Posterior computation:
  - Pr("GroundhogDay") $\propto$ Pr(Actor="Bill Murray") $\cdot$ Pr(ReleaseYear="1993") $\cdots$
  - Each Pr slot = value is computed in terms of LU outputs
- Soft KB-lookup: sample a movie according to the posterior
  - Randomization results in differentiability (similar to policy gradient alg.)
  - As opposed to using SQL queries to look up results deterministically
- Whole system can be trained using policy gradient & back-propagation

# Task-oriented Dialogue System



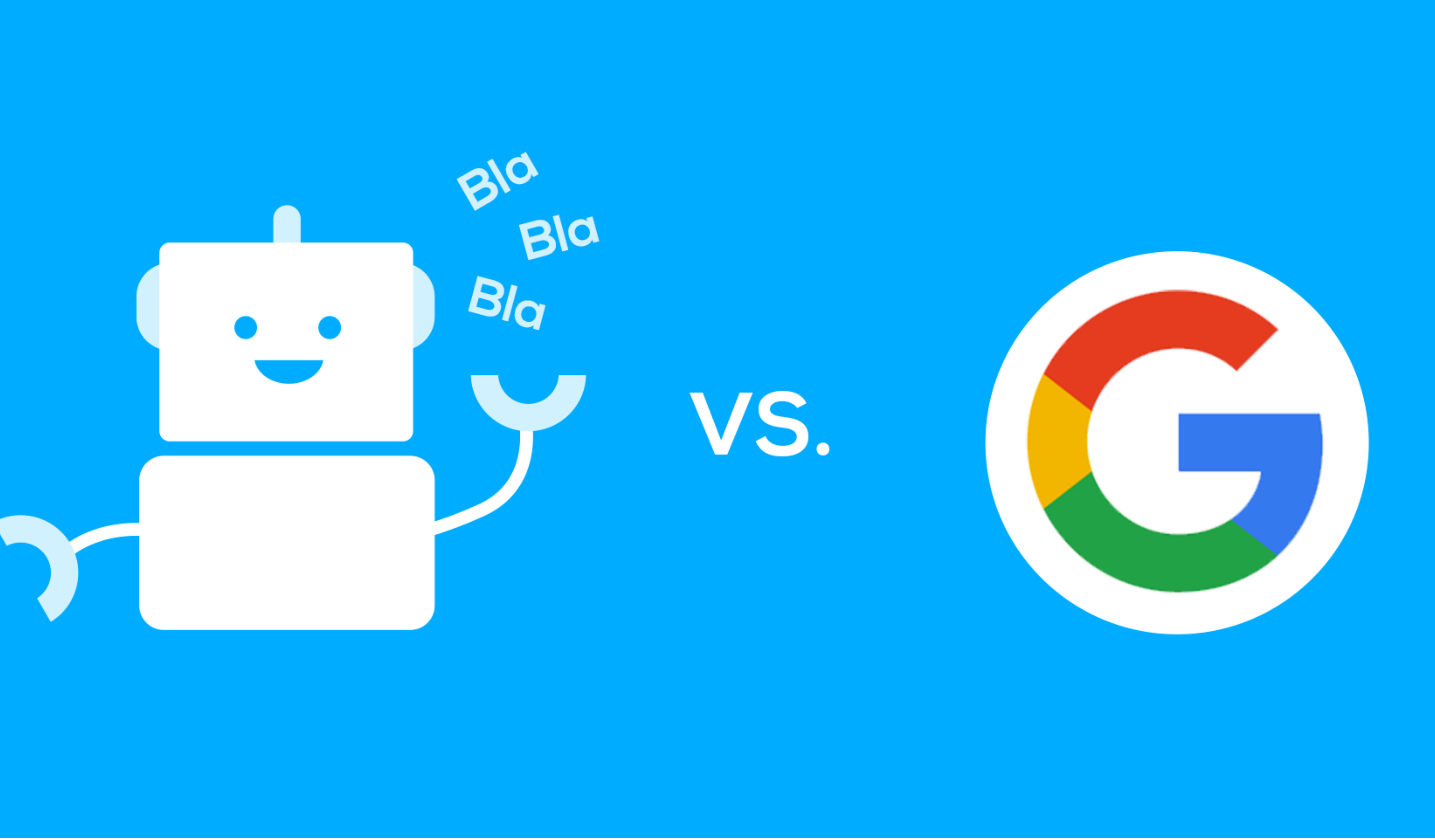

1. Use a single deep NN for {dialog manager and KB}

2. Recurrent network to track states of conversation

3. **Maintain (implicitly) a distribution over entities in KB**

4. A summary network to "summarize" distribution information

5. Multilayer perceptron policy network

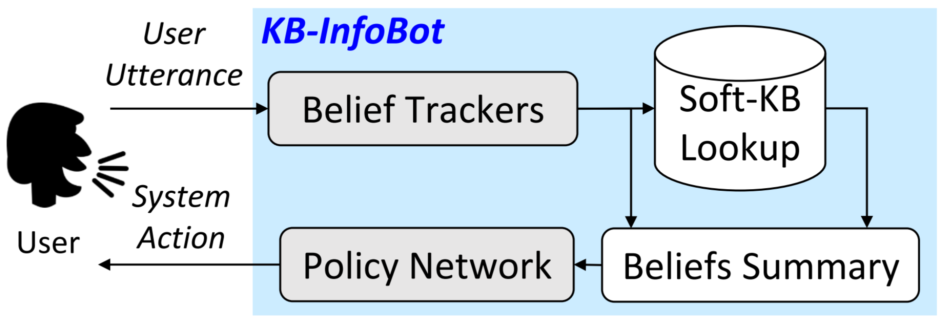



# Evaluation

for KB-InfoBot

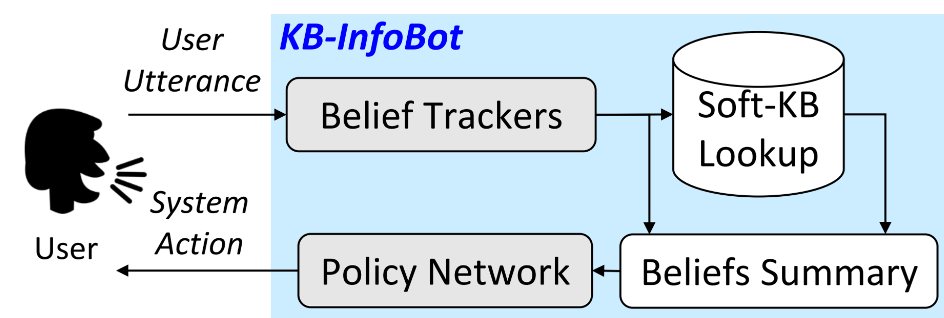
# Does Soft-KB lookup lead to better dialog policies?

- Belief Trackers:
  - A. Hand-Crafted (Bayesian updates)
  - B. Neural (GRU)
- Policy Network:
  - C. Hand-Crafted (Entropy Minimization)
  - D. Neural (GRU)
- KB-lookup:
  - 1. No KB lookup (Policy unaware of KB)
  - 2. Hard-KB lookup (SQL type lookup)
  - 3. Soft-KB lookup (KB Posterior)



# Does Soft-KB lookup lead to better dialog policies?

- Belief Trackers:
  - A. Hand-Crafted (Bayesian updates)
  - B. Neural (GRU)
- Policy Network:
  - C. Hand-Crafted (Entropy Minimization)
  - D. Neural (GRU)
- KB-lookup:
  - 1. No KB lookup (Policy unaware of KB)
  - 2. Hard-KB lookup (SQL type lookup)
  - 3. Soft-KB lookup (KB Posterior)




| | |
|---|---|
| **Rule-Based Agents:** | **A + C + (1, 2, 3)** |
| **RL-Based Agents:** | **A + D + (1, 2, 3)** |
| **E2E Agent:** | **B + D + (3)** |

# Does Soft-KB lookup lead to better dialog policies?

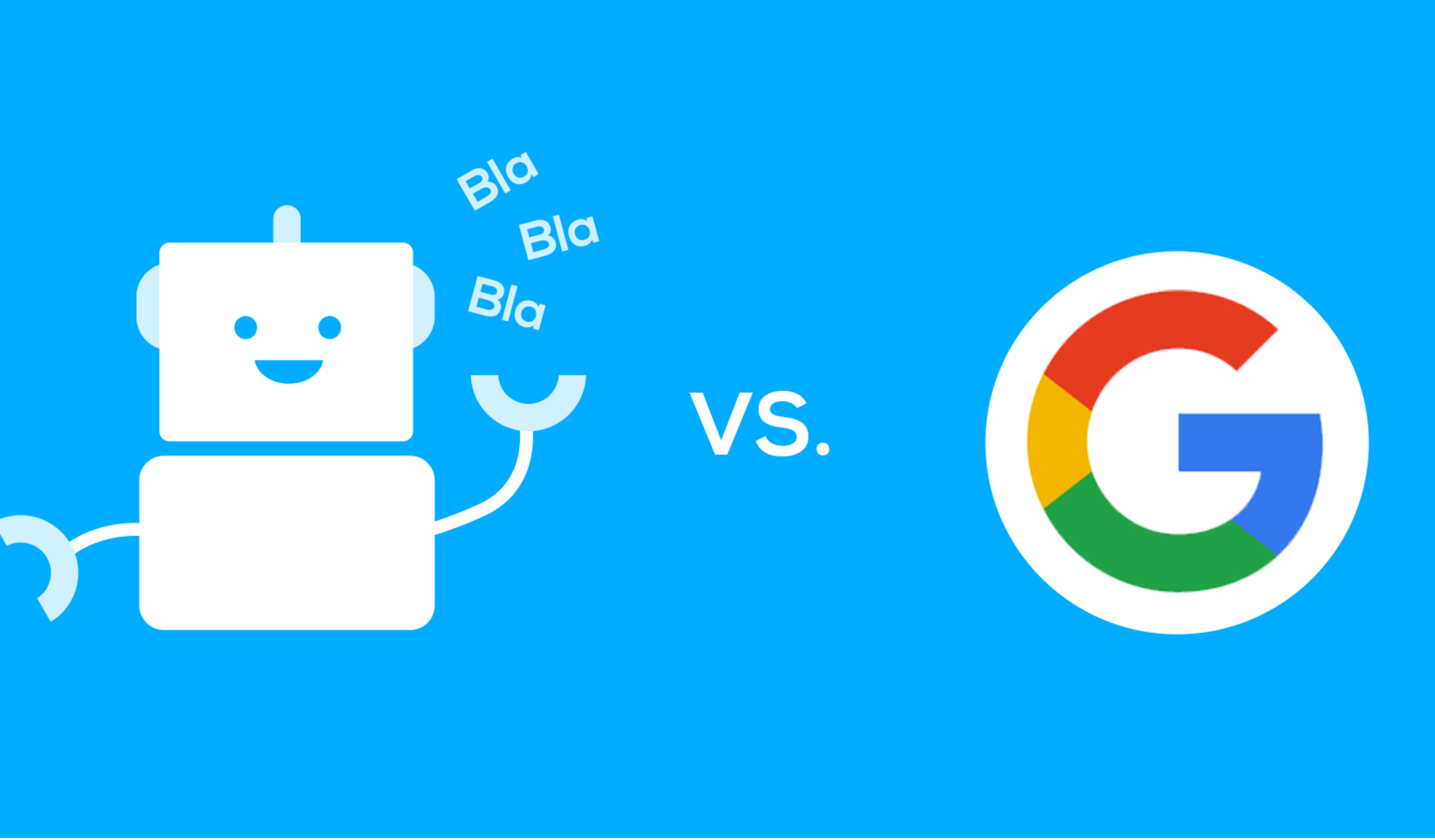
| | Agent | Small KB | | | Medium KB | | | Large KB | | | X-Large KB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T | S | R | T | S | R | T | S | R | T | S | R |
| No KB | Rule | 5.04 | .64 | .26±.02 | 5.05 | .77 | .74±.02 | 4.93 | .78 | .82±.02 | 4.84 | .66 | .43±.02 |
| | RL | 2.65 | .56 | .24±.02 | 3.32 | .76 | .87±.02 | 3.71 | .79 | .94±.02 | 3.64 | .64 | .50±.02 |
| Hard KB | Rule | 5.04 | .64 | .25±.02 | 3.66 | .73 | .75±.02 | 4.27 | .75 | .78±.02 | 4.84 | .65 | .42±.02 |
| | RL | 3.36 | .62 | .35±.02 | **3.07** | .75 | .86±.02 | 3.53 | .79 | .98±.02 | **2.88** | .62 | .53±.02 |
| Soft KB | Rule | **2.12** | .57 | .32±.02 | 3.94 | .76 | .83±.02 | 3.74 | .78 | .93±.02 | 4.51 | .66 | .51±.02 |
| | RL | 2.93 | .63 | .43±.02 | 3.37 | .80 | .98±.02 | 3.79 | .83 | 1.05±.02 | 3.65 | **.68** | **.62±.02** |
| | E2E | 3.13 | **.66** | **.48±.02** | 3.27 | **.83** | **1.10±.02** | **3.51** | **.83** | **1.10±.02** | 3.98 | .65 | .50±.02 |
| *Max* | | *3.44* | *1.0* | *1.64* | *2.96* | *1.0* | *1.78* | *3.26* | *1.0* | *1.73* | *3.97* | *1.0* | *1.37* |

- Metric:
  - # of Dialogue Turns (T)
  - Success Rate (correct movie returned) (S)
  - Average Reward (R)

**Rule-Based Agents:** A + C + (1, 2, 3)
**RL-Based Agents:** A + D + (1, 2, 3)
**E2E Agent:** B + D + (3)

# Conclusions

- Soft-KB lookup
    - Better dialogue policies
- E2E agent
    - Strong performance in simulations
    - Does not transfer to real interactions
    - Overfits to the limited natural language from the simulator
- Future research: personalized dialogue assistants?
    - Deploy using RL-Soft agent
    - Collect interactions to train E2E agent
    - Gradually switch to the E2E agent




Evaluation & Metrics

# Reward for RL ≈ Metric For Dialogue System

- Rating: correctness, appropriateness, and adequacy

| | |
|---|---|
| - Expert rating | high quality, high cost |
| - User rating | unreliable quality, medium cost |
| - Objective rating | Check desired aspects, low cost |

- Typical Reward Function
    - per turn penalty -1
    - Large reward at completion if successful
- e.g. KB-InfoBot
    - # of Dialogue Turns (T)
    - Success Rate (correct movie returned) (S)
    - Average Reward (R)

# Reward for RL ≈ Metric For Social Bots

- How NOT to use BLEU, ROUGE etc. [36]
- Instead good/bad, we measure responses from various aspects, e.g.,
    - Interestingness & Engagingness [37, 38]
    - Persona, consistency [39,40]
    - Contentfulness & usefulness [32]

# Engagingness - Ease of Answering

- forward-looking function: the constraints a turn places on the next turn [38]

utterance with a dull response. We manually constructed a list of dull responses $\mathbb{S}$ consisting 8 turns such as "I don't know what you are talking about", "I have no idea", etc., that we and others have found occur very frequently in SEQ2SEQ models of conversations. The reward function is given as follows:

$$r_1 = -\frac{1}{N_{\mathbb{S}}} \sum_{s \in \mathbb{S}} \frac{1}{N_s} \log p_{\text{seq2seq}}(s|a) \qquad (1)$$

where $N_{\mathbb{S}}$ denotes the cardinality of $N_{\mathbb{S}}$ and $N_s$ denotes the number of tokens in the dull response $s$.

# Challenges of RL

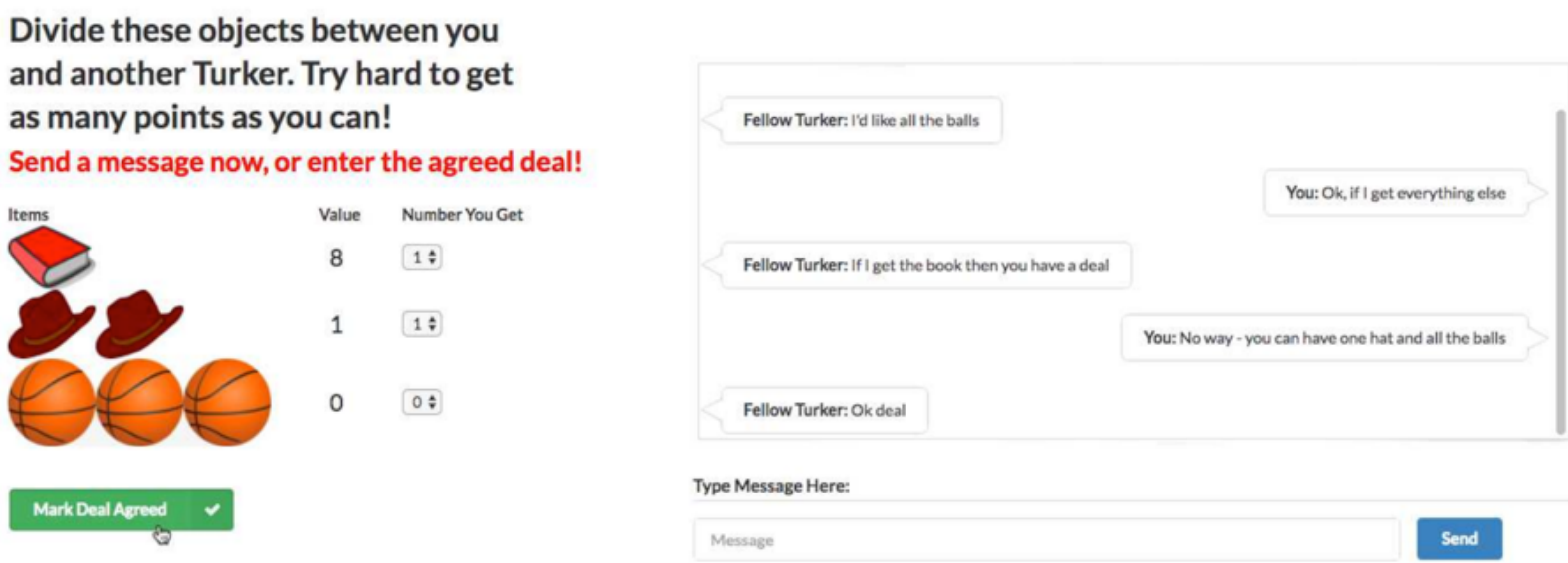
- Complex, (unbounded) state-action space
- Evaluation feedback, (delayed) **reward**
- Non-stationarity
- Need for trial and error, to explore as well as exploit
    - how an agent can learn from success and failure, from reward and punishment
    - one constantly has to decide btw continuing in a comfortable existence and striking out into unknown in the hopes of discovering a new and better life.

# Challenges of RL

- Complex, (unbounded) state-action space
- Evaluation feedback, (delayed) **reward — sparse, implicit, inaccurate**
- Non-stationarity
- Need for trial and error, to explore as well as exploit
  - how an agent can learn from success and failure, from reward and punishment
  - one constantly has to decide btw continuing in a comfortable existence and striking out into unknown in the hopes of discovering a new and better life.

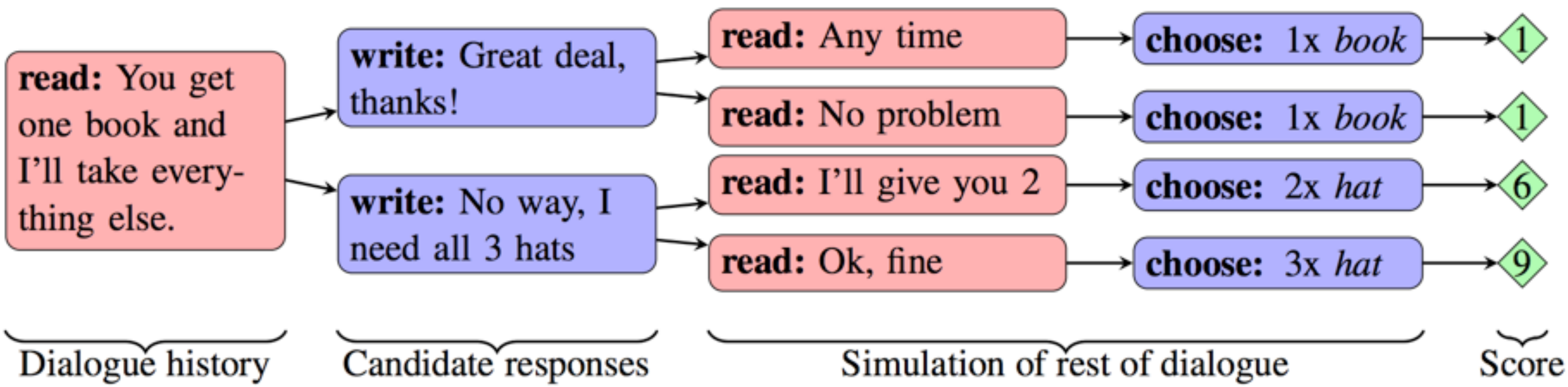
# Special Setting

- [29]



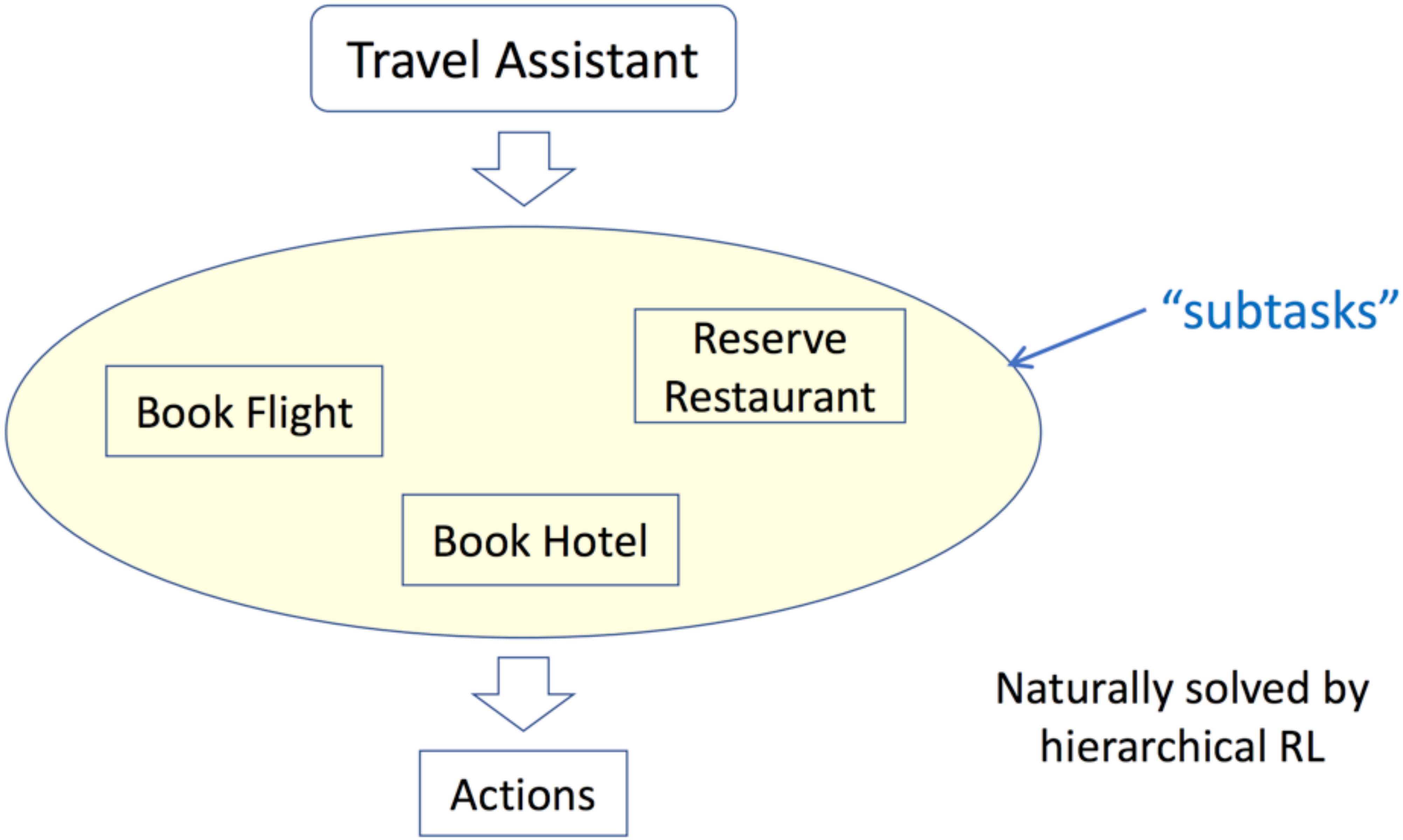
# Special Setting

- [29]

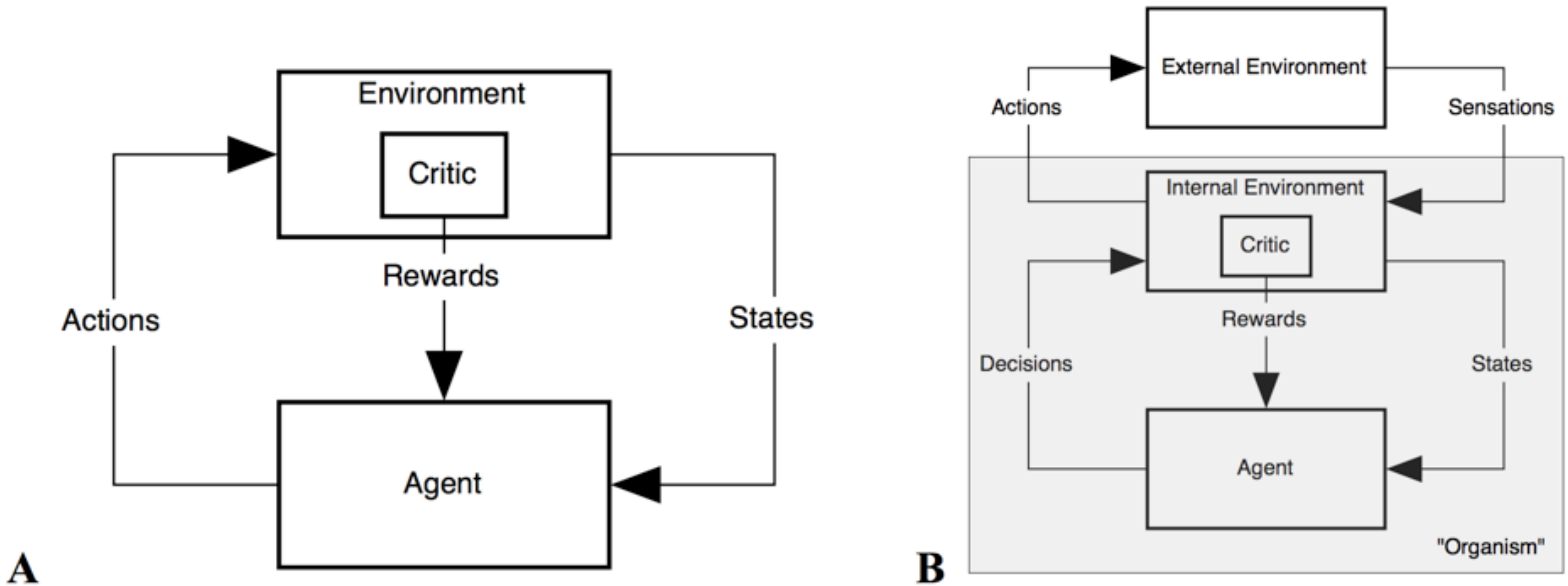


# Composite Task in Dialogue

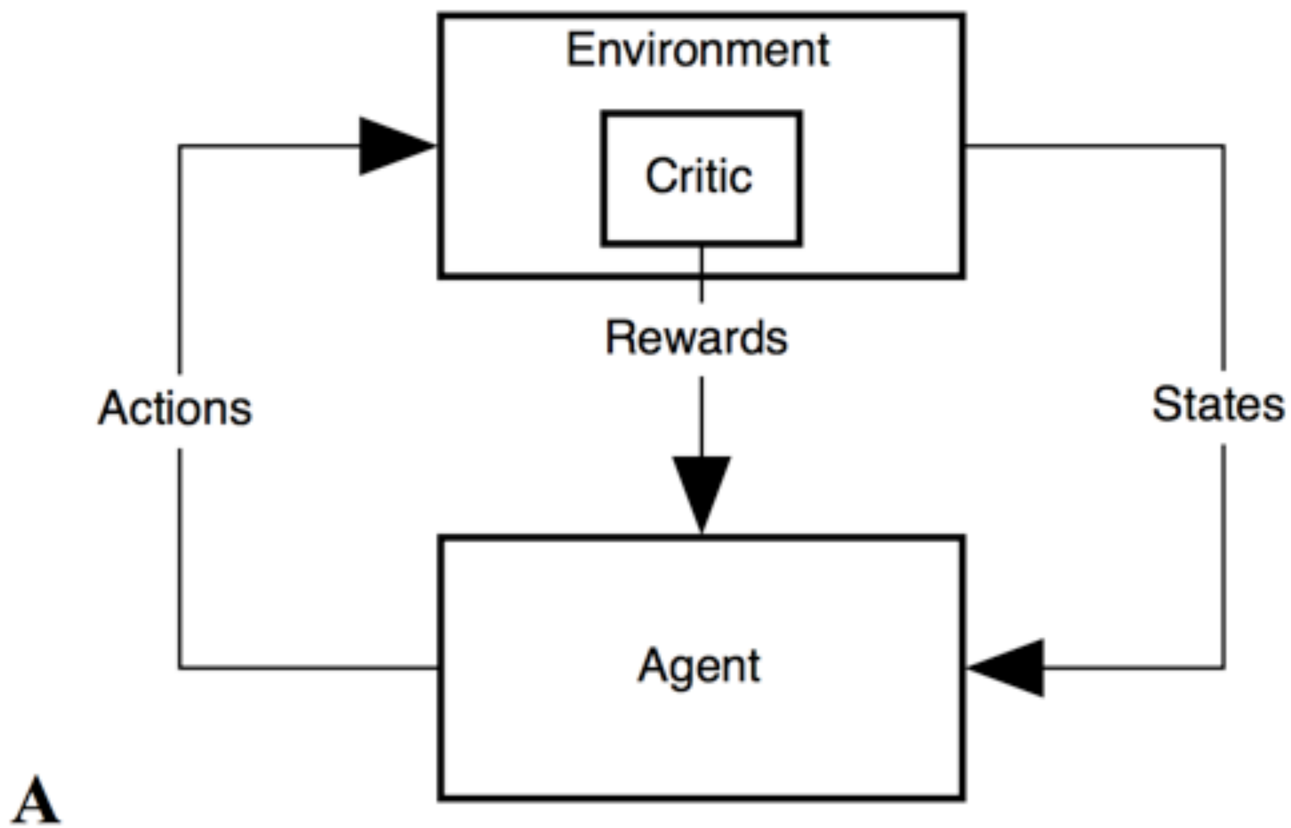
- [35]

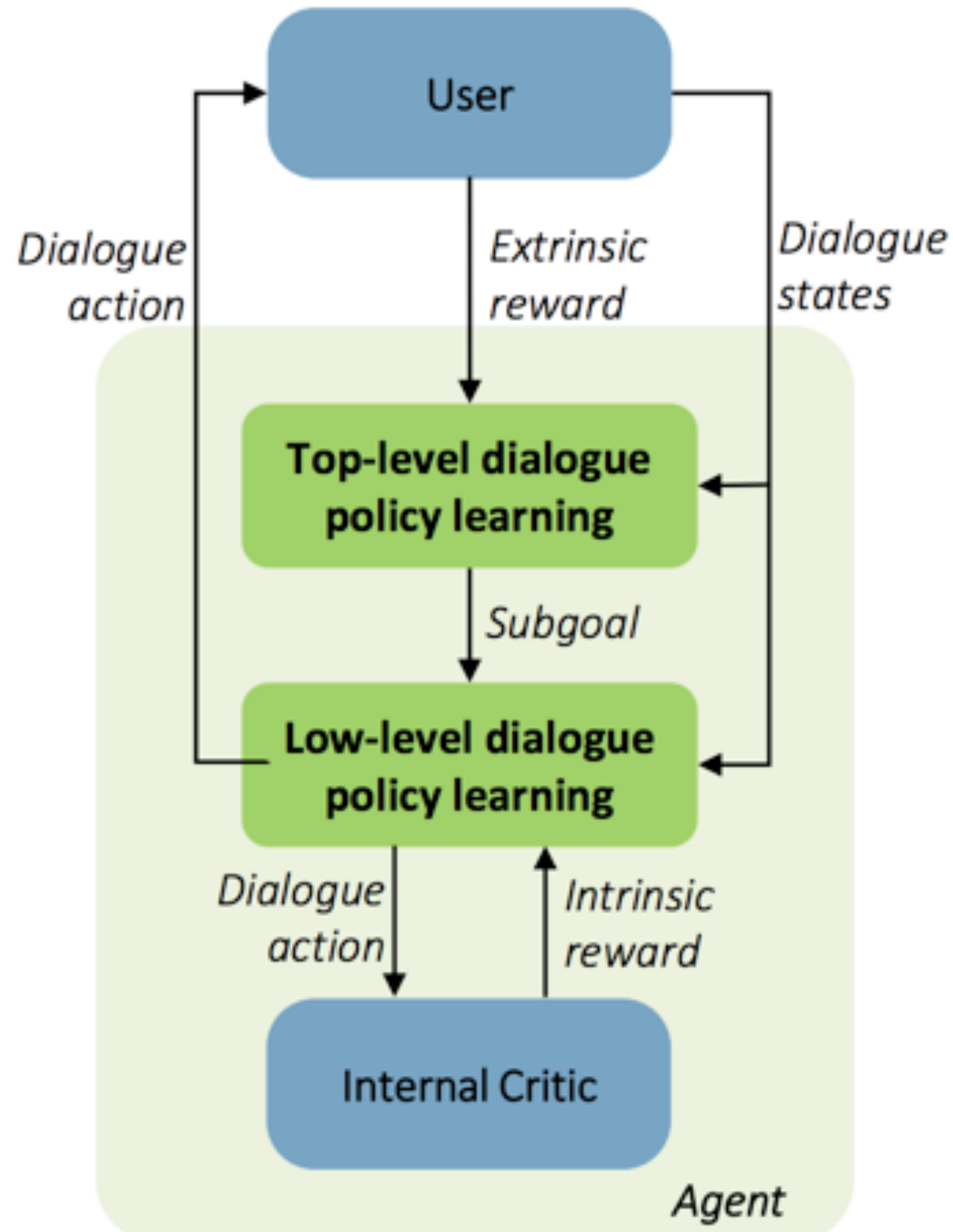


# External Reward V.S. Internal Reward



# External Reward V.S. Internal Reward



Thanks for your attention!

Q&A

# References


- [1] Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio. "On the difficulty of training recurrent neural networks." International Conference on Machine Learning. 2013.
- [2] Mikolov, Tomas, et al. "Recurrent neural network based language model." Interspeech. Vol. 2. 2010.
- [3] Colah. “Understanding LSTMs”. http://colah.github.io/posts/2015-08-Understanding-LSTMs/
- [4] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.
- [5] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).
- [6] Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).
- [7] Wen, Tsung-Hsien, et al. "A network-based end-to-end trainable task-oriented dialogue system." arXiv preprint arXiv:1604.04562 (2016).
- [8] Bowman, Samuel R., et al. "Generating sentences from a continuous space." arXiv preprint arXiv:1511.06349 (2015).
- [9] Hu, Baotian, et al. "Convolutional neural network architectures for matching natural language sentences." Advances in neural information processing systems. 2014.

# References


- [10] Qiu, Xipeng, and Xuanjing Huang. "Convolutional Neural Tensor Network Architecture for Community-Based Question Answering." IJCAI. 2015.
- [11] Wu, Yu, et al. "Sequential Match Network: A New Architecture for Multi-turn Response Selection in Retrieval-based Chatbots." ACL 2017.
- [12] https://github.com/seatgeek/fuzzywuzzy
- [13] https://www.linkedin.com/pulse/duplicate-quora-question-abhishek-thakur
- [14] Is That a Duplicate Quora Question? https://www.linkedin.com/pulse/duplicate-quora-question-abhishek-thakur
- [15] https://github.com/HouJP/kaggle-quora-question-pairs/
- [16] https://pan.baidu.com/s/1dEV01gd
- [17] Lifeng Shang, Zhengdong Lu, Hang Li. “Neural Responding Machine for Short-Text Conversation”. EMNLP 2015.
- [18] Serban, Iulian Vlad, et al. "A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues." AAAI. 2017.
- [19] Serban, Iulian Vlad, et al. "Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation." AAAI. 2017.
- [20] Alessandro Sordoni, Yoshua Bengio et al., “A Hierarchical Recurrent Encoder-Decoder for Context-Aware Generative Query Suggestion”. CIKM 2015 slides.

# References


- [21] Zhou Hao et al, “Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory”. arXiv preprint 2017.
- [22] Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, Guoping Long. “A Conditional Variational Framework for Dialog Generation”. ACL 2017.
- [23] Jessica Ficler, Yoav Goldberg. “Controlling Linguistic Style Aspects in Neural Language Generation”. arXiv preprint 2017.
- [24] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, Sune Lehmann. “Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm”. EMNLP 2017.
- [25] DeepMoji. https://deepmoji.mit.edu/
- [26] Louis Shao, et al. “Generating Long and Diverse Responses with Neural Conversation Models”. arXiv preprint 2017.
- [27] Iulian Vlad Serban, et al. “Generating Factoid QuestionsWith Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus”. ACL 2016a.
- [28] Yun-Nung Chen et al. “End-to-End Memory Networks with Knowledge Carryover for Multi-Turn Spoken Language Understanding”. arXiv preprint 2016b.
- [29] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, Dhruv Batra. “Deal or No Deal? End-to-End Learning for Negotiation Dialogues”. EMNLP 2017.

# References


- [30] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, Li Deng. “Towards End-to-End Reinforcement Learning of Dialogue Agents for Information Access”. ACL 2017.
- [31] Chen Xing et al. “Topic Aware Neural Response Generation”. arXiv preprint 2016.
- [32] Marjan Ghazvininejad et al. “A Knowledge-Grounded Neural Conversation Model”. arXiv preprint 2017.
- [33] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, Dhruv Batra. “Visual Dialog”. CVPR 2017.
- [34] Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, Dhruv Batra. “Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning”. arXiv preprint 2017.
- [35]  Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, Kam-Fai Wong. “Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning”. EMNLP 2017.
- [36] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, Joelle Pineau. “How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation”. EMNLP 2016.
- [37] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, Bill Dolan. “A Diversity-Promoting Objective Function for Neural Conversation Models”. NAACL 2016.
- [38] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao and Dan Jurafsky. “Deep Reinforcement Learning for Dialogue Generation”. EMNLP 2016.
- [39] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, Bill Dolan. “A Persona-Based Neural Conversation Model”. ACL 2016.
- [40] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, Xiaoyan Zhu. “Assigning personality/identity to a chatting machine for coherent conversation generation”. arXiv preprint 2017.