# COMPANION TEACHING

## Towards Affordable On-line Dialogue Policy Learning


## ASAF 2017

Runzhe Yang @ SJTU ACM CLASS

# A Brief Self Introduction

Runzhe Yang @ SJTU ACM CLASS

# About me:

**Runzhe Yang** (Chinese: 杨闰哲),
3rd-year undergrad at ACM Class
Research intern at SJTU Speech Lab
mentored by Prof. **Kai Yu**
on **Spoken Dialogue Systems.**

# About my mentor:

## Prof. Kai Yu (Chinese: 俞凯),

**Research Professor,** Dept. of CSE, SJTU

Co-founder & Chief Scientist of AISPEECH

B.Eng & M.Sc from Tsinghua Univ.

Ph.D. from MIL, **Cambridge**

Senior member of IEEE, Member of ISCA

Member of IEEE Speech and Language Processing Technical Committee

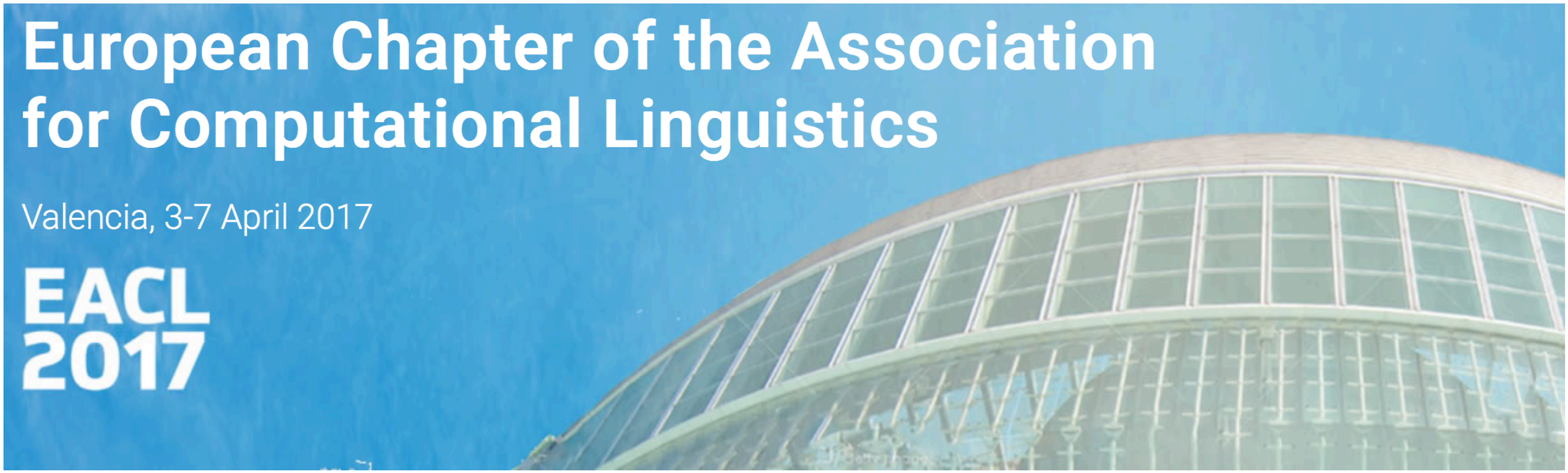
**1000 Overseas Talent Plan (Young Talent)**

# About SpeechLab:


**“To carry out speech-related research, you must be an excellent engineer; however, to become an outstanding engineer, you must be a good scientist.”**

Runzhe Yang @ SJTU ACM CLASS

# About this paper:




- Currently once every three years
- Short papers: 120/504 (24% acceptance rate)
  - Compare to 46/199 in 2014

# COMPANION TEACHING

## Towards Affordable On-line Dialogue Policy Learning

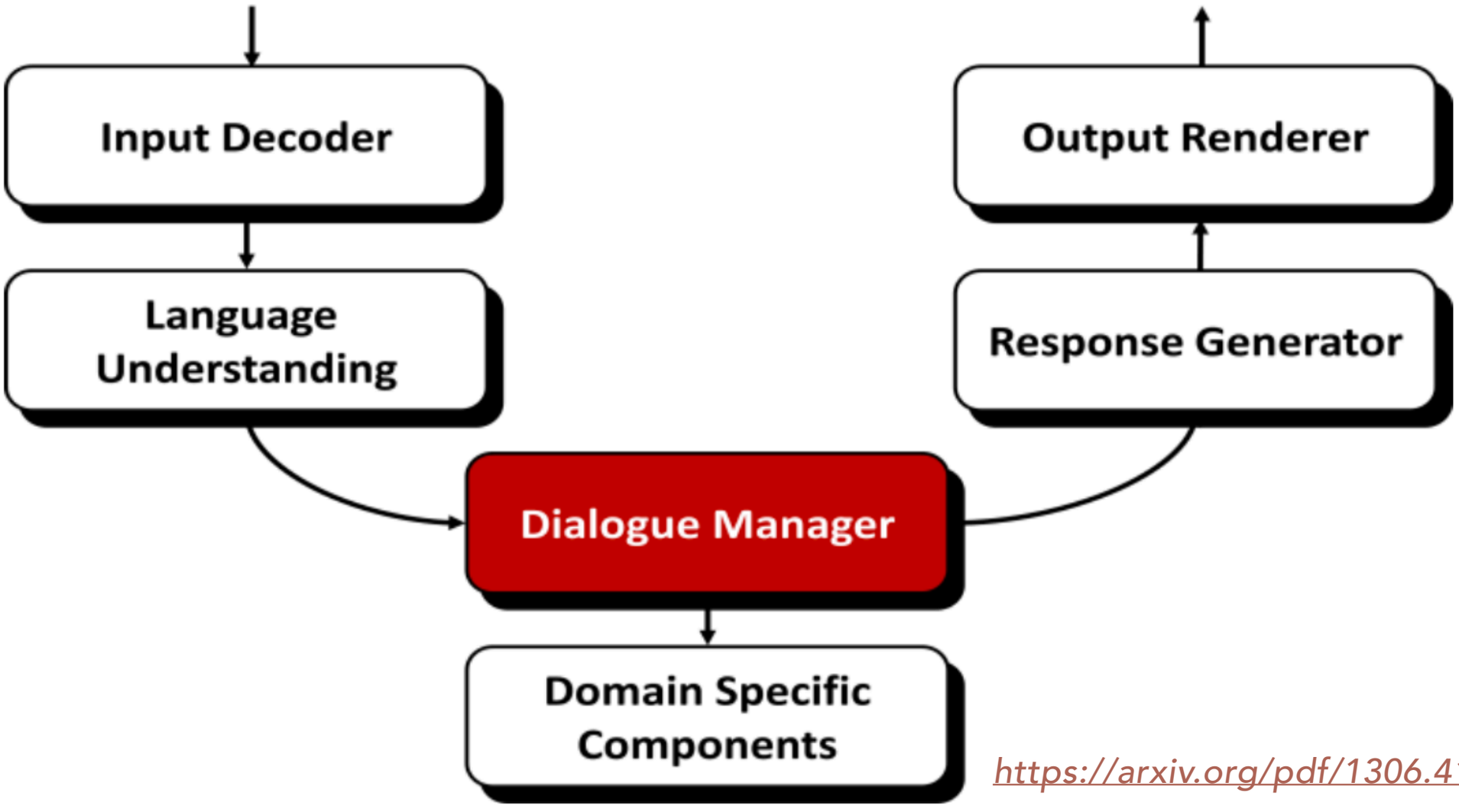

## ASAF 2017

Runzhe Yang @ SJTU ACM CLASS

# On-line Dialogue Policy Learning?

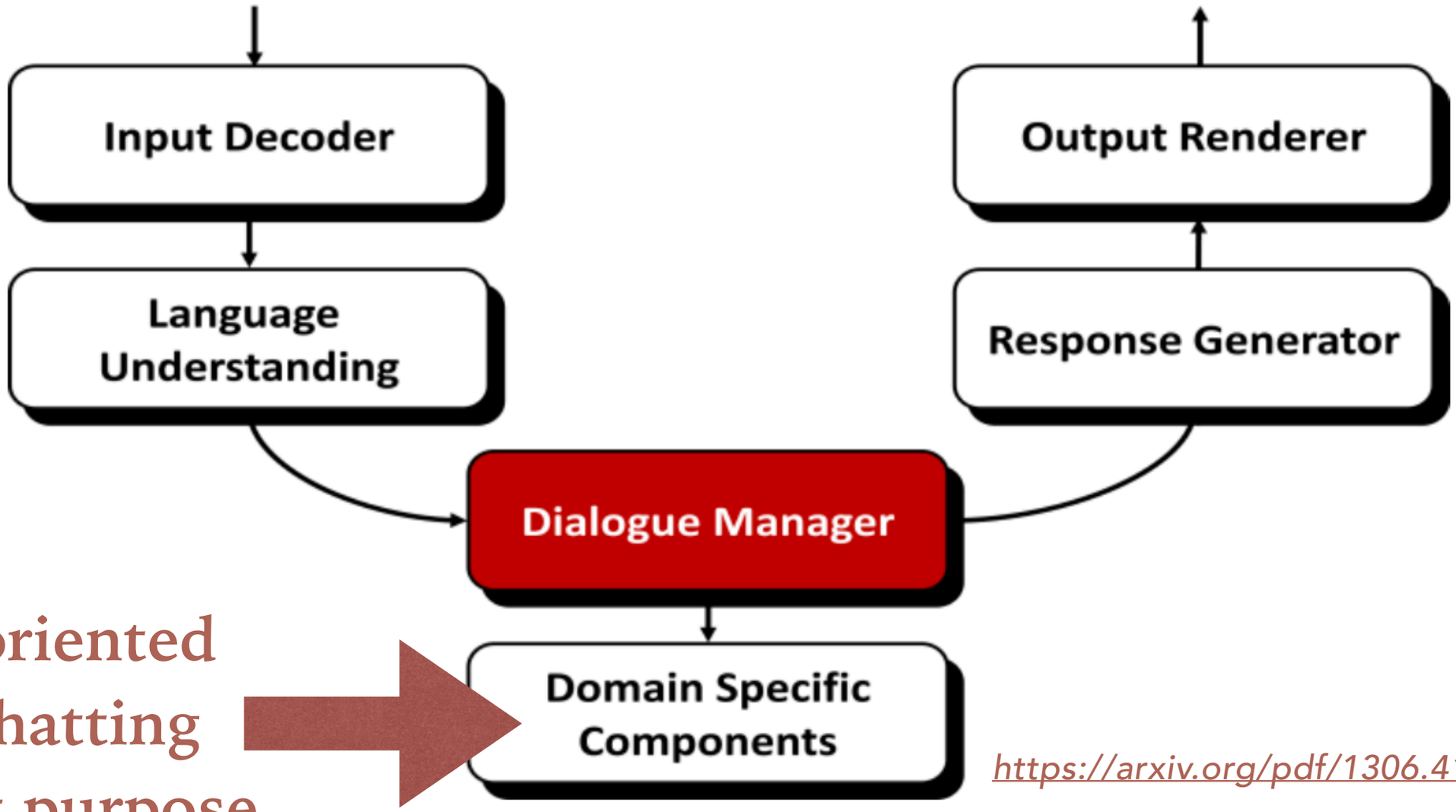
Runzhe Yang @ SJTU ACM CLASS

# What's a dialogue system?




https://arxiv.org/pdf/1306.4134.pdf

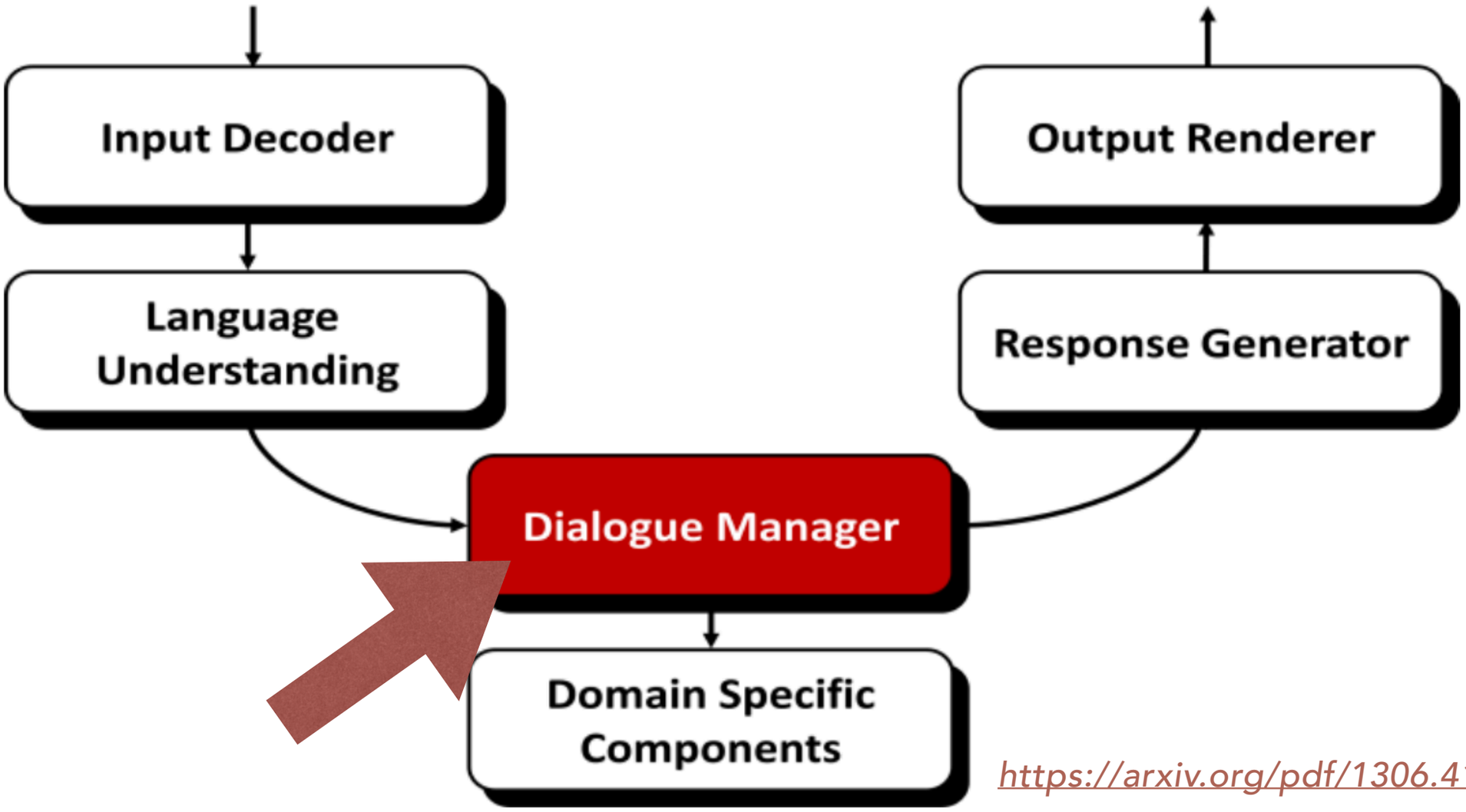
# What's a dialogue system?




Task-oriented
Not chatting
without purpose

https://arxiv.org/pdf/1306.4134.pdf

# What's a dialogue system?

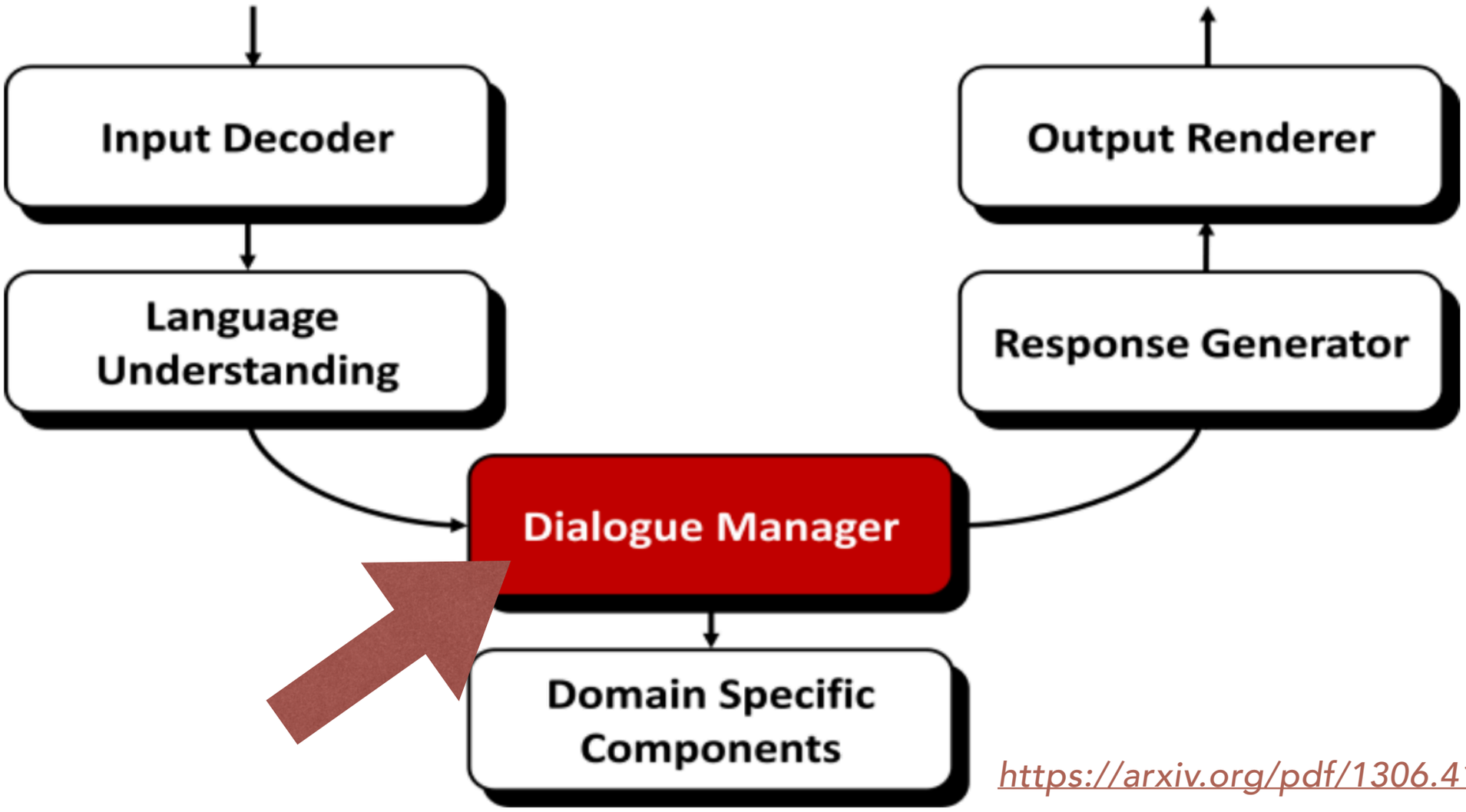



https://arxiv.org/pdf/1306.4134.pdf

# Example of Successful Dialogue

| TASK: ask for *italian* restaurant in *north* area & request its *phone number* | | | | | | |
|---|---|---|---|---|---|---|
| | | Dialogue Turn | Score | $Q^{turn}$ | $Q^{succ}$ | FP |
| [1] | System | [SLU] welcomemsg() | | | | |
| | User | [Top ASR] Italian food in the north part of town. | 0.30 | -4.54 | 27.44 | False |
| [2] | System | [SLU] expl-conf(food="italian") | | | | |
| | User | [Top ASR] Yes. | 0.99 | -2.24 | 29.09 | False |
| [3] | System | [SLU] offer(name="caffe uno") inform(food="italian") inform(area="north") | | | | |
| | User | [Top ASR] The phone number. | 0.92 | -2.00 | 28.27 | False |
| [4] | System | [SLU] offer(name="caffe uno") inform(food="italian") inform(area="north") Inform(phone="01223314954") | | | | |
| | User | [Top ASR] Does it serve danish italian food. | 0.53 | -2.41 | 28.20 | False |
| [5] | System | [SLU] offer(name="caffe uno") inform(food="italian") inform(area="north") | | | | |
| | User | [Top ASR] Goodbye. | 0.58 | 0.05 | 27.42 | False |

# What's a dialogue system?




https://arxiv.org/pdf/1306.4134.pdf

# How do we build this “brain”?

# How do we build this “brain”?


- **rule-based methods**
  - hand-craft rules, “safe” but not “flexible”

# How do we build this “brain”?


- **rule-based methods**
  - hand-craft rules, “safe” but not “flexible”
- **data-driven methods**
  - learn from data / interactions!

# How do we build this “brain”?


- **data-driven methods**
  - learn from data / interactions!

# How do we build this “brain”?

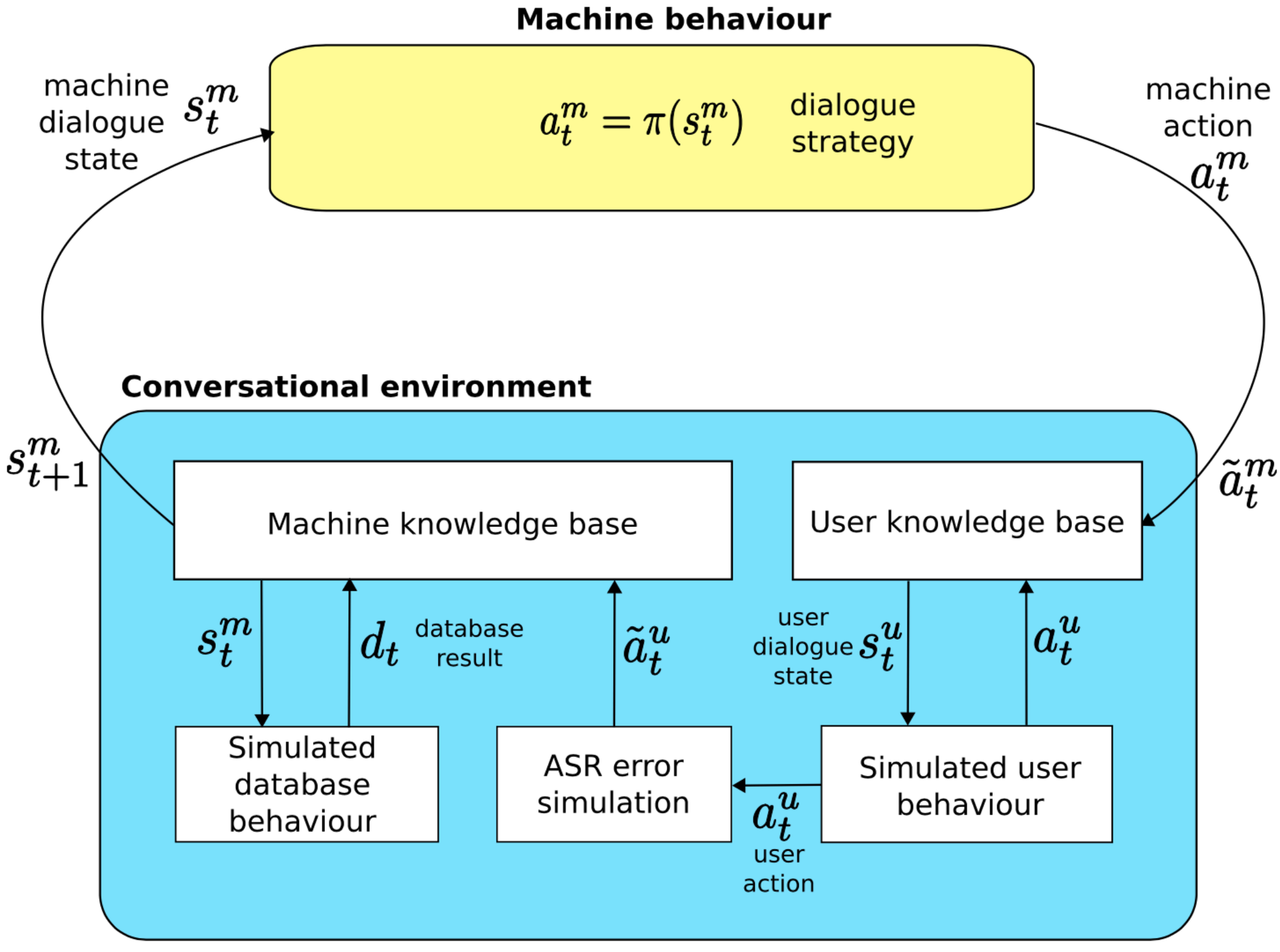

- **data-driven methods**
  - learn from data / interactions!
  - Partially Observable Markov Decision Process (POMDP)

# How do we build this “brain”?


- data-driven methods
  - learn from data / interactions!
  - Partially Observable Markov Decision Process (POMDP)
  - <S, A, T, R, Ω, O, γ>

# How do we build this “brain”?


- data-driven methods
  - learn from data / interactions!
  - Partially Observable Markov Decision Process (POMDP)
  - <S, A, T, R, Ω, O, γ>
  - Reinforcement Learning: π(a|o) maximize total reward




http://www.cstr.ed.ac.uk/downloads/publications/2009/cuayahuitl-csl09.pdf

# How do we build this “brain”?


- **Where can we get data / interactions ?**

# How do we build this “brain”?


- **Where can we get data / interactions ?**
- task oriented dialogue data is rare…

# How do we build this “brain”?


- **Where can we get data / interactions ?**

- task oriented dialogue data is rare…
- train the system with “user simulators”

# How do we build this “brain”?


- **Where can we get data / interactions ?**
- task oriented dialogue data is rare…
- train the system with “user simulators”
  - or paid testers

- **Where can we get data / interactions ?**
  - task oriented dialogue data is rare…
  - train the system with “user simulators”
    - or paid testers
- **Reality: non-cooperative!**

- **Where can we get data / interactions ?**
  - task oriented dialogue data is rare…
  - train the system with “user simulators”
    - or paid testers
  - **Reality: non-cooperative!**
  - **Not evolvable!**

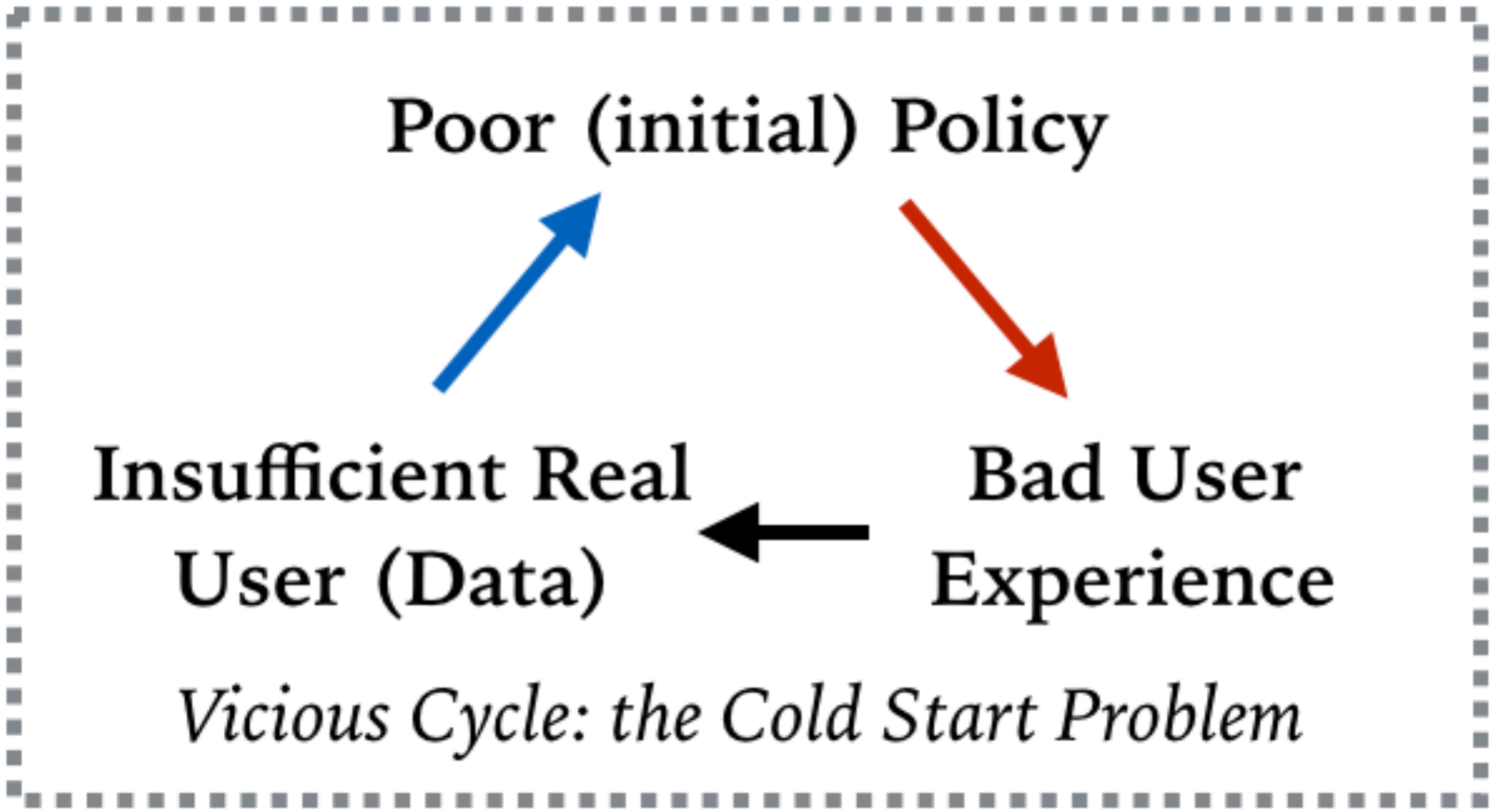
- **Where can we get data / interactions ?**
  - task oriented dialogue data is rare…
  - train the system with “user simulators”
    - or paid testers

- **Reality: non-cooperative!**
- **Not evolvable!**
- **Must be on-line!**

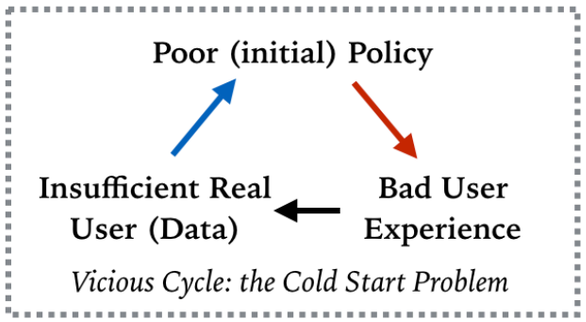
# How to Build Evolvable Conversational Agent in Real World Scenarios?

Runzhe Yang @ SJTU ACM CLASS

# How to Build Evolvable Conversational Agent in Real World Scenarios?



# How to Build Evolvable Conversational Agent in Real World Scenarios?




## **Possible Solutions** to break the vicious cycle

- Inefficient Learning Process (Solvable) ✓
- Unsafe Policy Behavior (Solvable) ✓
- Individual Rationality (Unsolvable) ✗

**Therefore, an ideal on-line policy learning framework should be measured using following two criteria:**

→ Inefficient Learning Process (Solvable) ✓

→ Unsafe Policy Behavior (Solvable) ✓

**Therefore, an ideal on-line policy learning framework should be measured using following two criteria:**

→ Inefficient Learning Process (Solvable) ✓

***Efficiency*** reflects how long it takes for the on-line policy learning algorithm to reach a satisfactory performance level.

→ Unsafe Policy Behavior (Solvable) ✓

**Therefore, an ideal on-line policy learning framework should be measured using following two criteria:**

→ Inefficient Learning Process (Solvable) ✓

***Efficiency*** reflects how long it takes for the on-line policy learning algorithm to reach a satisfactory performance level.

→ Unsafe Policy Behavior (Solvable) ✓

***Safety**** reflects whether the initial policy can satisfy the quality-of-service requirement in real-world scenarios during on-line policy learning period.

**Therefore, an ideal on-line policy learning framework should be measured using following two criteria:**

→ Inefficient Learning Process (Solvable) ✓

→ Unsafe Policy Behavior (Solvable) ✓

* Most previous studies of on-line policy learning have been focused on the ***efficiency*** **issue**, such as
  - Gaussian process reinforcement learning (GPRL)(Gasic et al., 2010),
  - Deep reinforcement learning (DRL) (Fatemi et al., 2016; Williams and Zweig, 2016; Su et al., 2016), etc.

**Therefore, an ideal on-line policy learning framework should be measured using following two criteria:**

→ Inefficient Learning Process (Solvable) ✓

→ Unsafe Policy Behavior (Solvable) ✓

* However, ***safety*** **is a prerequisite for the efficiency to be achieved**.
  - **Reason**: an unsafe on-line learned policy can consequently fail to attract sufficient real users to continuously improve the policy, no matter how efficient the algorithm is.
  - **Urgency**: on the ***safety*** **issue** which little work has been done.

# Our Solution: Human-in-the-loop

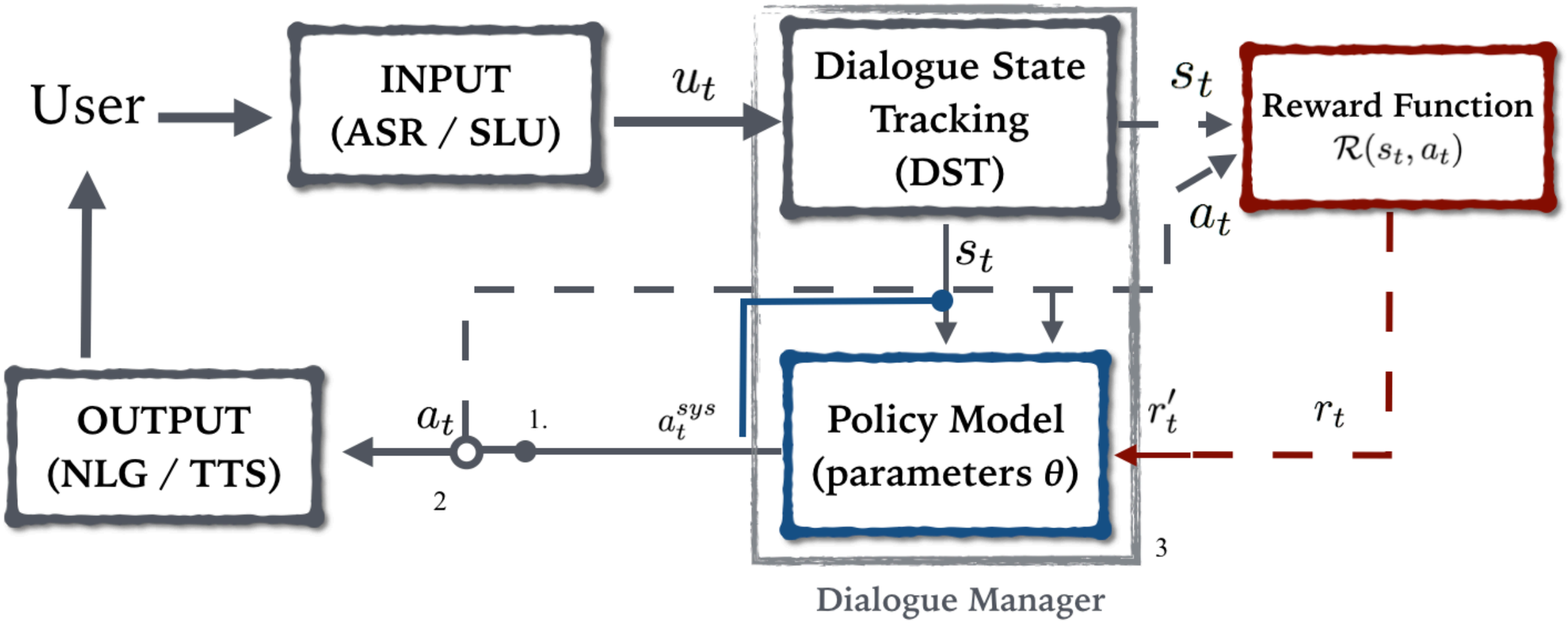
# Our Solution: Human-in-the-loop

In this work, we propose a **human-machine hybrid RL framework**, ***Companion Teaching***, which includes a human teacher in the on-line dialogue policy training loop. The involved human teacher accompanies the machine and provides **immediate hands-on guidance at turn level** during on-line policy learning period. This will lead to a ***safer*** policy learning process since the learning is done before any possible dialogue failure at the end.

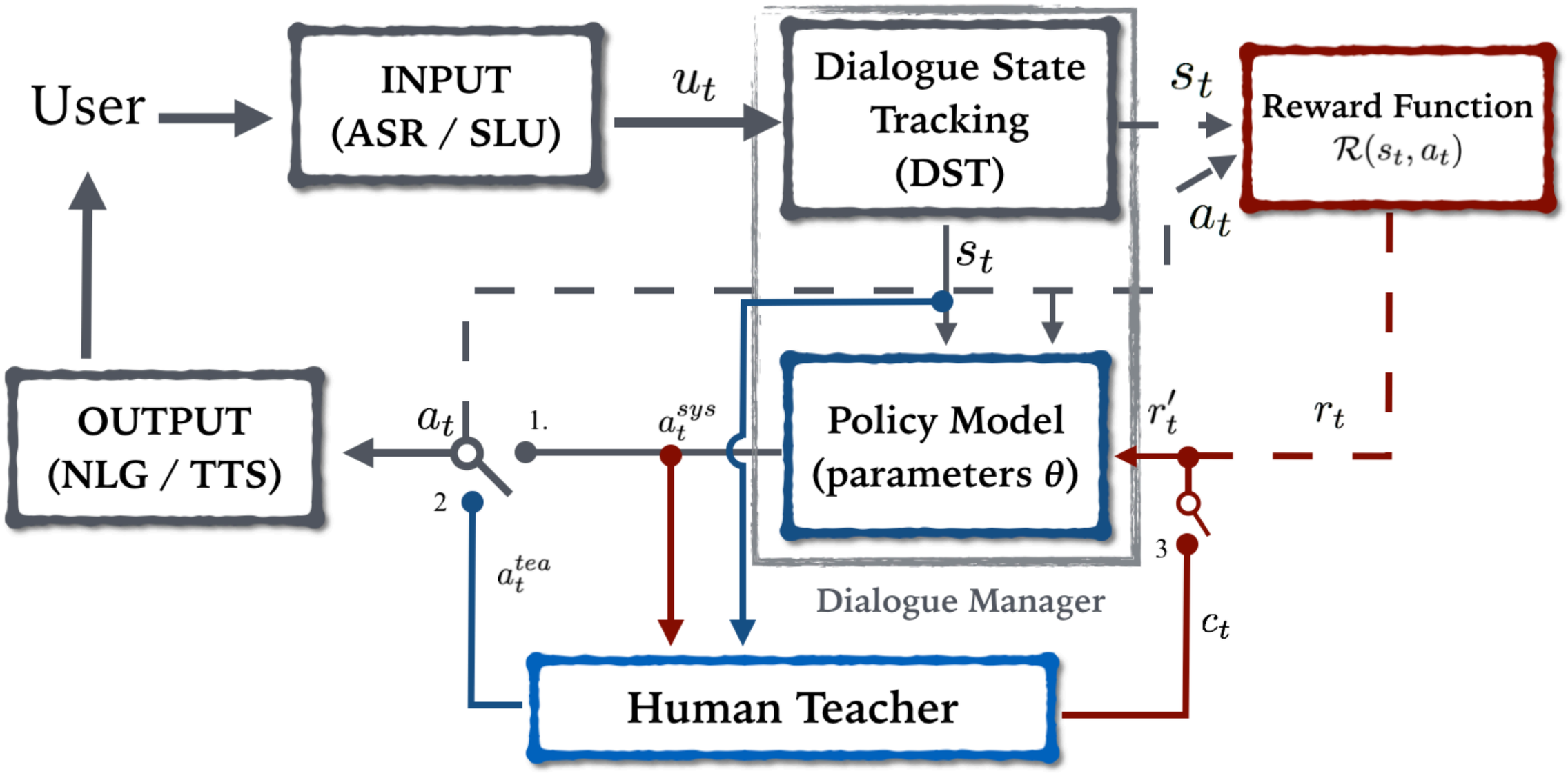
# Our Solution: Human-in-the-loop

In this work, we propose a **human-machine hybrid RL framework**, ***Companion Teaching***, which includes a human teacher in the on-line dialogue policy training loop. The involved human teacher accompanies the machine and provides **immediate hands-on guidance at turn level** during on-line policy learning period. This will lead to a ***safer*** policy learning process since the learning is done before any possible dialogue failure at the end.

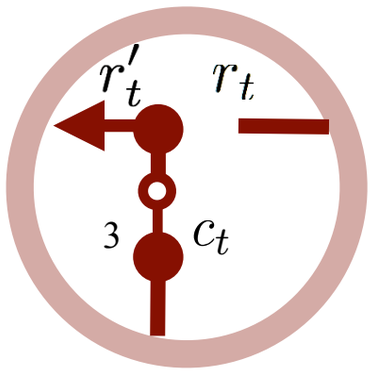
# RL-Based Framework

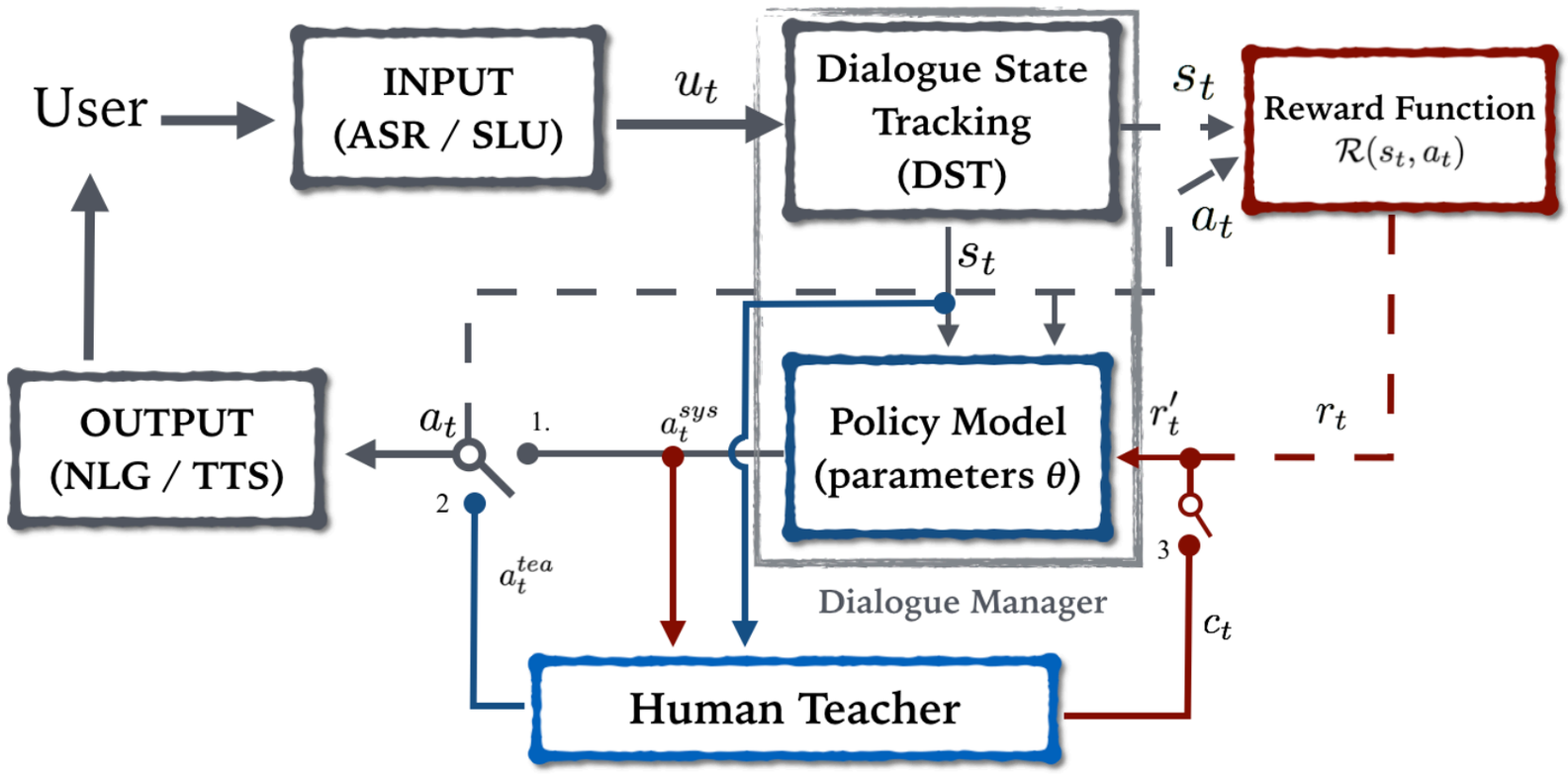


# Companion Teaching Framework



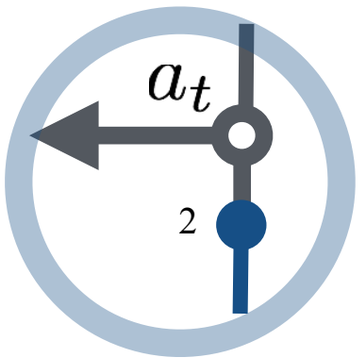
# Teaching Strategies

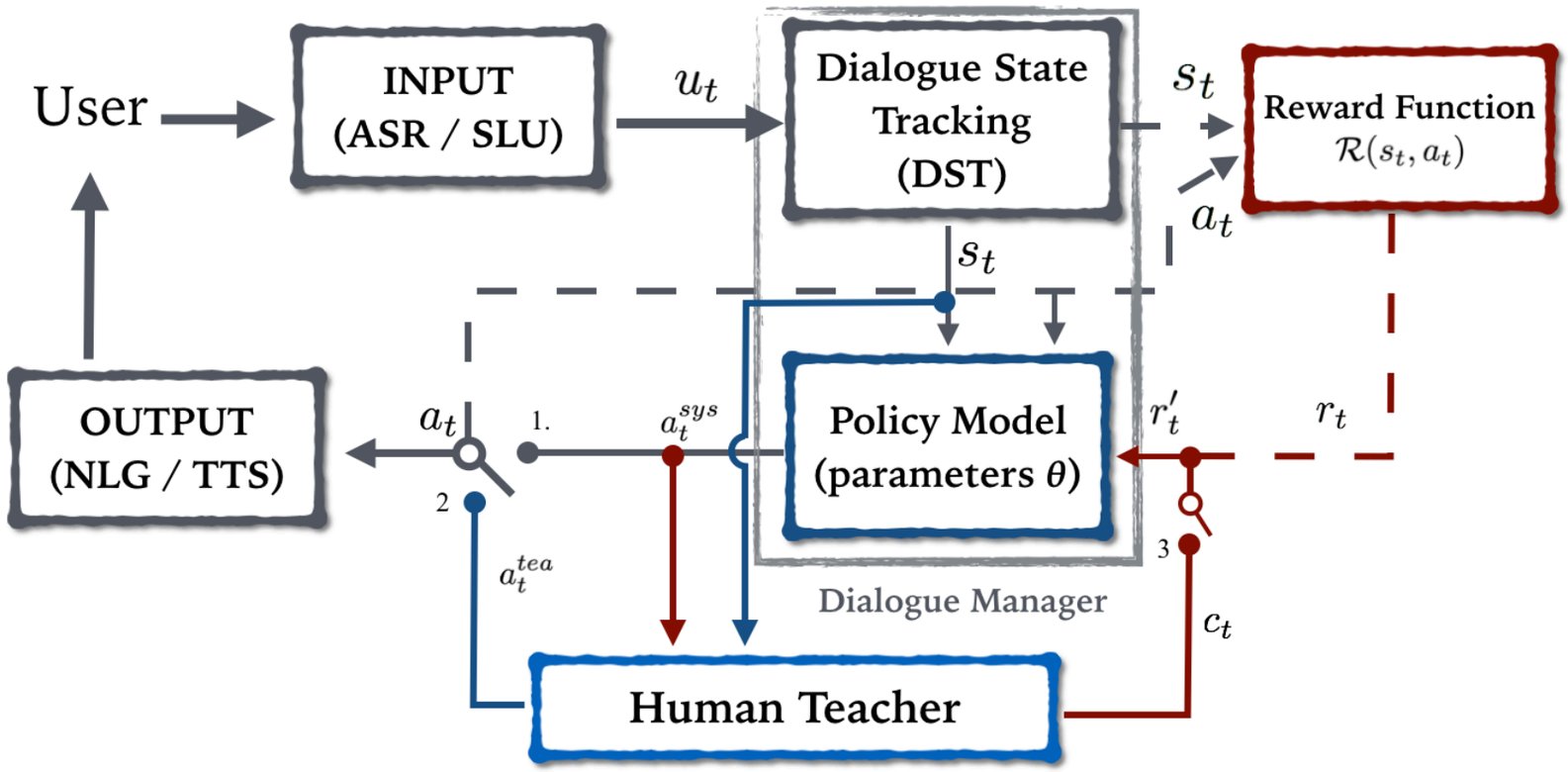



**Teaching via Critic Advice (CA)** corresponds to the **right switch (position 3)** in Figure 1. The key idea is to give the policy model an ***extra immediate reward signal*** from teacher, which differentiates between good actions and bad actions.

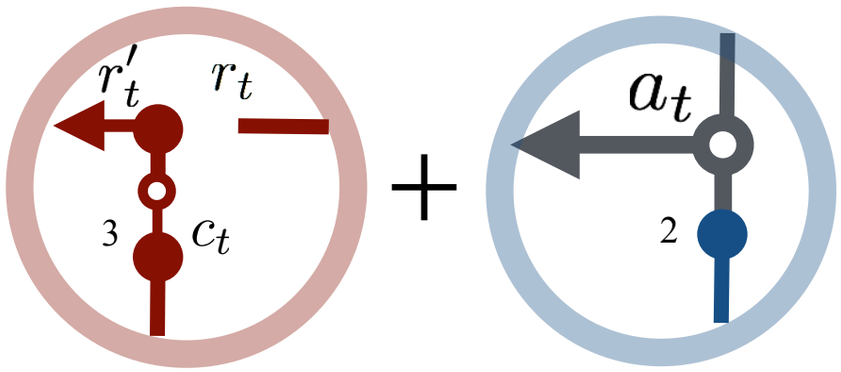
# Teaching Strategies

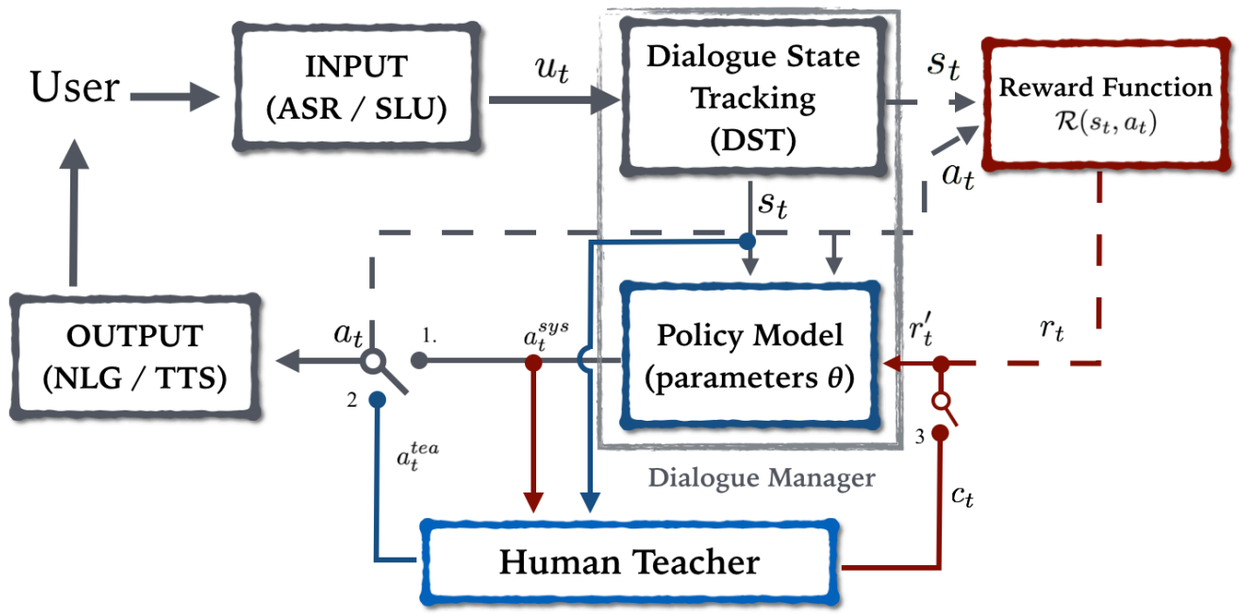



**Teaching via Example Action (EA)** corresponds to the **left switch (position 2)**. The human teacher ***directly gives an example action*** at a particular state. The system can learn from teacher's action by considering the action as its own exploration action.

# Teaching Strategies




**Teaching via Example Action with Predicted Critique (EAPC)** take advantages of both EA and CA. The human teacher ***gives an example action*** and meanwhile, an ***extra reward*** will be given to the policy model as well. And this extra reward signal lasts even in teacher's absence. The example actions will be collected to train a ***weak action prediction model***.

# Experiments & Results

# Experiments & Results

- *Dataset:* Dialogue State Tracking Challenge 2 (DSTC2) dataset

# Experiments & Results

- *Dataset:* Dialogue State Tracking Challenge 2 (DSTC2) dataset
- *DST:* a Rule-based Tracker (Sun et al., 2014)

# Experiments & Results

- *Dataset:* Dialogue State Tracking Challenge 2 (DSTC2) dataset
- *DST:* a Rule-based Tracker (Sun et al., 2014)
- ***Policy Model:*** a Deep Q-Network (DQN) (Mnih et al., 2015)
  - Two hidden layers to map a belief state $s_t$ to the values of the possible actions $a_t$ at that state, $Q(s_t, a_t; \theta)$.
  - a target network with weight vector $\theta^-$ is used.

# Experiments & Results

- *Dataset:* Dialogue State Tracking Challenge 2 (DSTC2) dataset
- *DST:* a Rule-based Tracker (Sun et al., 2014)
- ***Policy Model:*** a Deep Q-Network (DQN) (Mnih et al., 2015)
  - Two hidden layers to map a belief state $s_t$ to the values of the possible actions $a_t$ at that state, $Q(s_t, a_t; \theta)$.
  - a target network with weight vector $\theta^-$ is used.
- ***Reward Design:*** consisting of three parts
  - Length penalty: *-1* at each turn;
  - Success bonus: *+30* at the end of the session;
  - Extra reward: $1 \leq c_t \leq 20$.

# Experiments & Results

- *User Simulator: an agenda-based user simulator (Schatzmann et al., 2007)*

# Experiments & Results

- *User Simulator: an agenda-based user simulator (Schatzmann et al., 2007)*
- ***Simulated Teacher:*** a **well-trained policy model** with success rate 0.78 in our experiment.

# Experiments & Results

- *User Simulator: an agenda-based user simulator (Schatzmann et al., 2007)*
- ***Simulated Teacher:*** a **well-trained policy model** with success rate 0.78 in our experiment.
- ***Teaching Budget:*** 1500 turns

# Experiments & Results

- *User Simulator: an agenda-based user simulator (Schatzmann et al., 2007)*
- ***Simulated Teacher:*** a **well-trained policy model** with success rate 0.78 in our experiment.
- ***Teaching Budget:*** 1500 turns

**Evaluating Safety:**
The moving success rate-#dialogue curve in training in which the real performance experienced by users.

# Experiments & Results

- *User Simulator: an agenda-based user simulator (Schatzmann et al., 2007)*
- ***Simulated Teacher:*** a **well-trained policy model** with success rate 0.78 in our experiment.
- ***Teaching Budget:*** 1500 turns

**Evaluating Safety:**
The moving success rate-#dialogue curve in training in which the real performance experienced by users.

**Evaluating efficiency:**
It can be evaluated by the number of dialogues required to achieve a reasonable performance in the testing curve.

# Experiments & Results

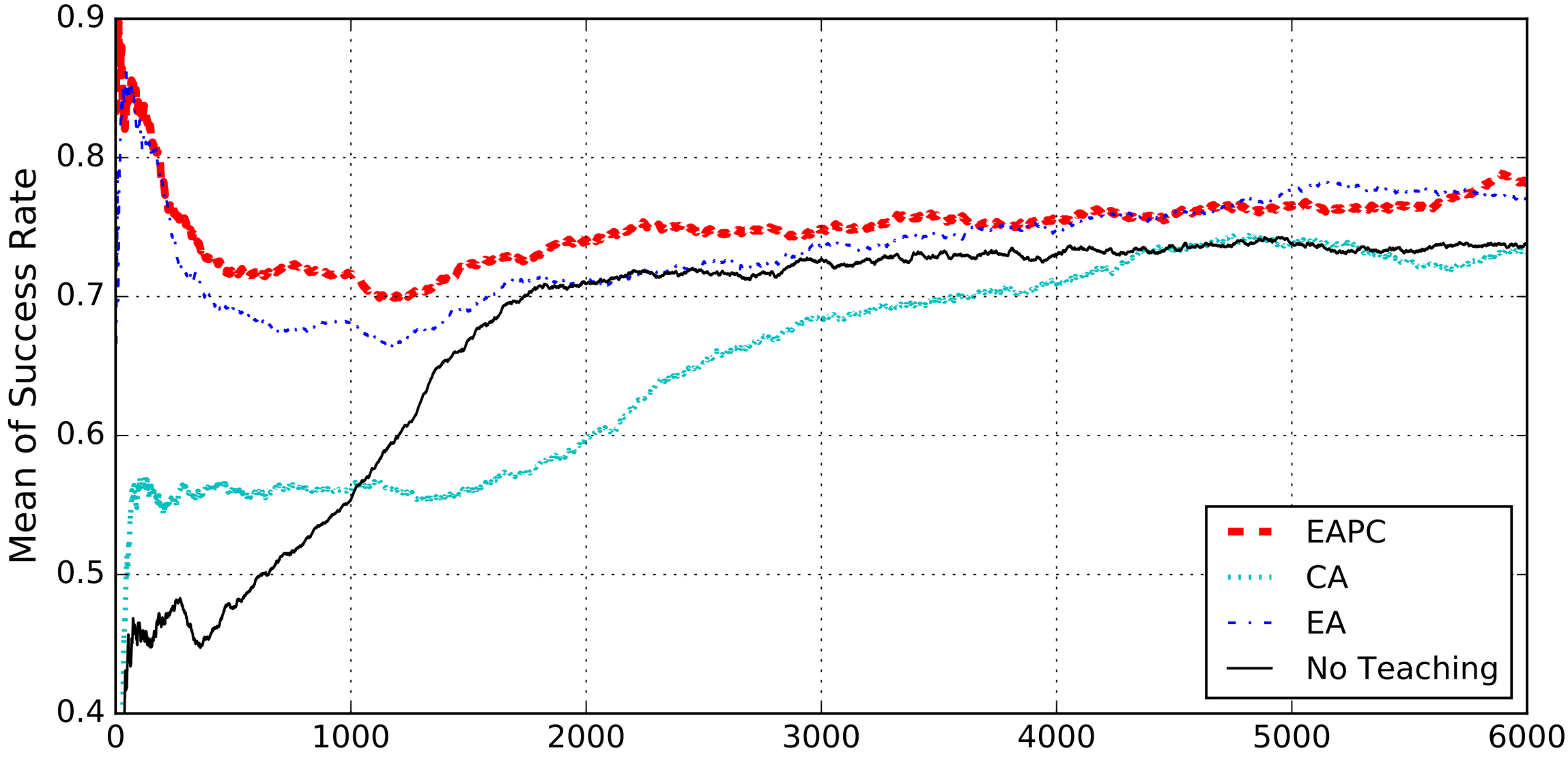



Figure 2: The training curves of moving average success rate.

# Experiments & Results

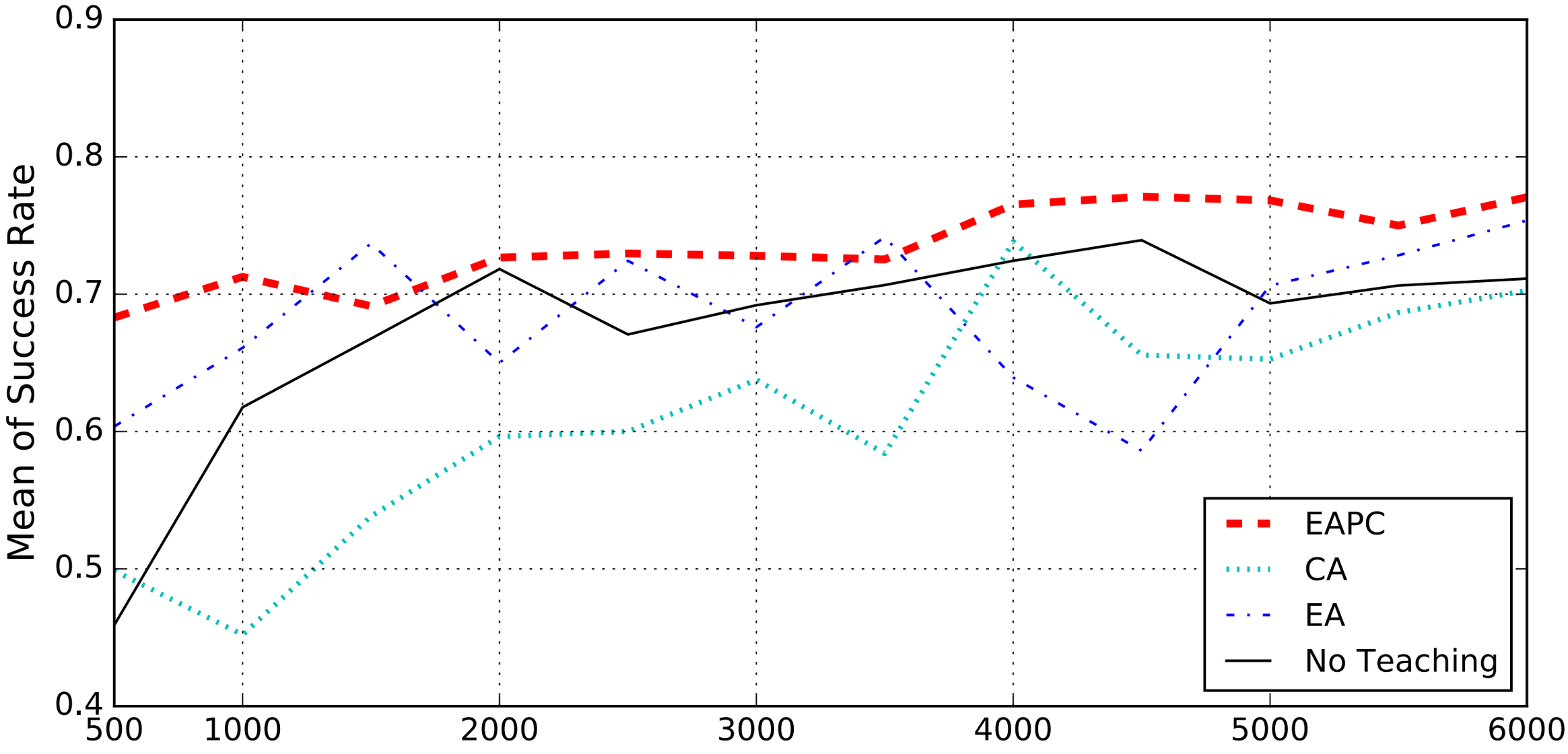



Figure 3: The testing curves of moving average success rate.

# Discussion & Future work

- Timing: when to teach?

# Discussion & Future work

- Timing: when to teach?
- Quantitive evaluation
- Hitting time & Risk index

# Discussion & Future work

- Timing: when to teach?
- Quantitive evaluation
  - Hitting time & Risk index
- Rule-based system as guidance

# Discussion & Future work

- Timing: when to teach?
- Quantitive evaluation
  - Hitting time & Risk index
- Rule-based system as guidance
- Experiments in the real conversational environment

# Thanks!

Runzhe Yang @ SJTU ACM CLASS