EVALUATING CONVERSATIONAL RECOMMENDER SYSTEMS VIA USER SIMULATION

Shuo Zhang*
Bloomberg, London, UK
@imsure318

Krisztian Balog
University of Stavanger, Norway
@krisztianbalog

* Work done while at the University of Stavanger, Norway.

Bloomberg
Engineering

Information Access
& Interaction
http://iai.group

MOTIVATION

- Test-collection based ("offline") evaluation
  - Possible to create a reusable test collection for a specific subtask ✓
  - Limited to a single turn, does not measure overall user satisfaction ✗
- Human evaluation
  - Possible to annotate entire conversations ✓
  - Expensive, time-consuming, does not scale ✗
- Evaluation of conversational information access systems is an open challenge. We explore **user simulation** in this work.

OBJECTIVES

- Develop a user simulator that
  - produces responses that a real user would give in a certain dialog situation
  - enables automatic assessment of conversational agents
  - makes no assumptions about the inner workings of conversational agents
  - is data-driven (requires only a small annotated dialogue corpus)

PROBLEM STATEMENT

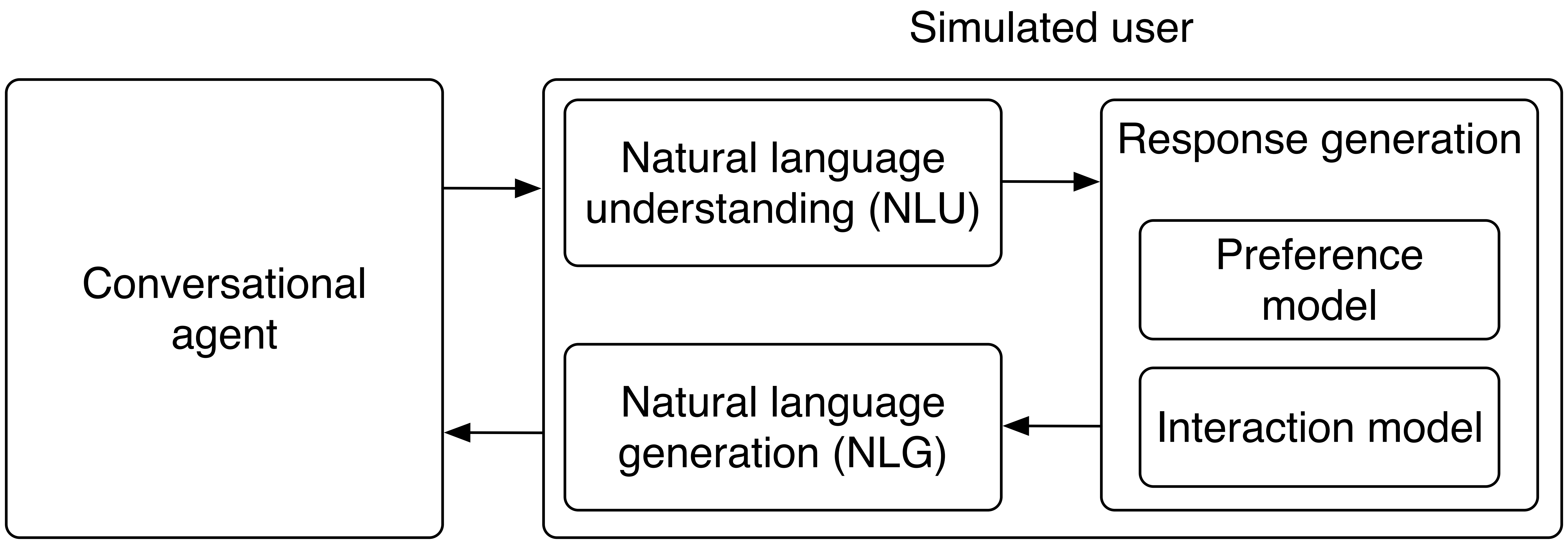
- For a given system S and user population U, the goal of user simulation U^* is to predict the performance of S when used by U, denoted as $M(S,U)$
- For two systems S_1 and S_2, U^* should be such that

$$\text{if} \quad M(S_1,U) < M(S_2,U)$$

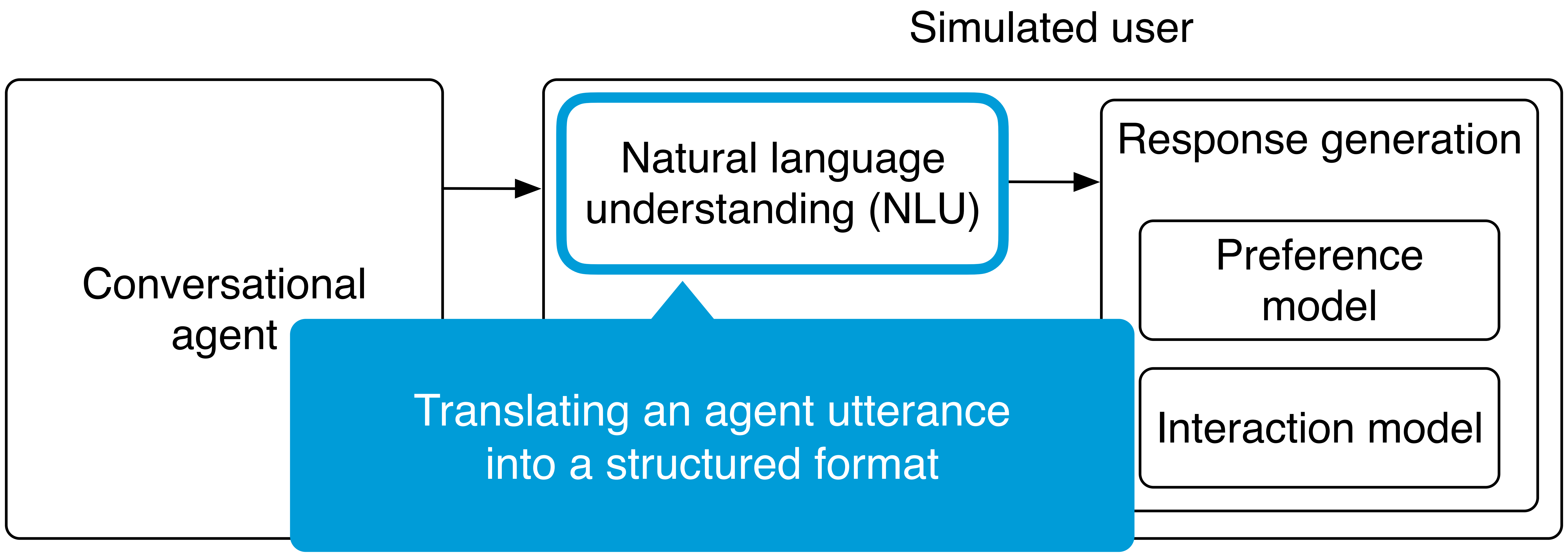
$$\text{then} \quad M(S_1,U^*) < M(S_2,U^*)$$

APPROACH

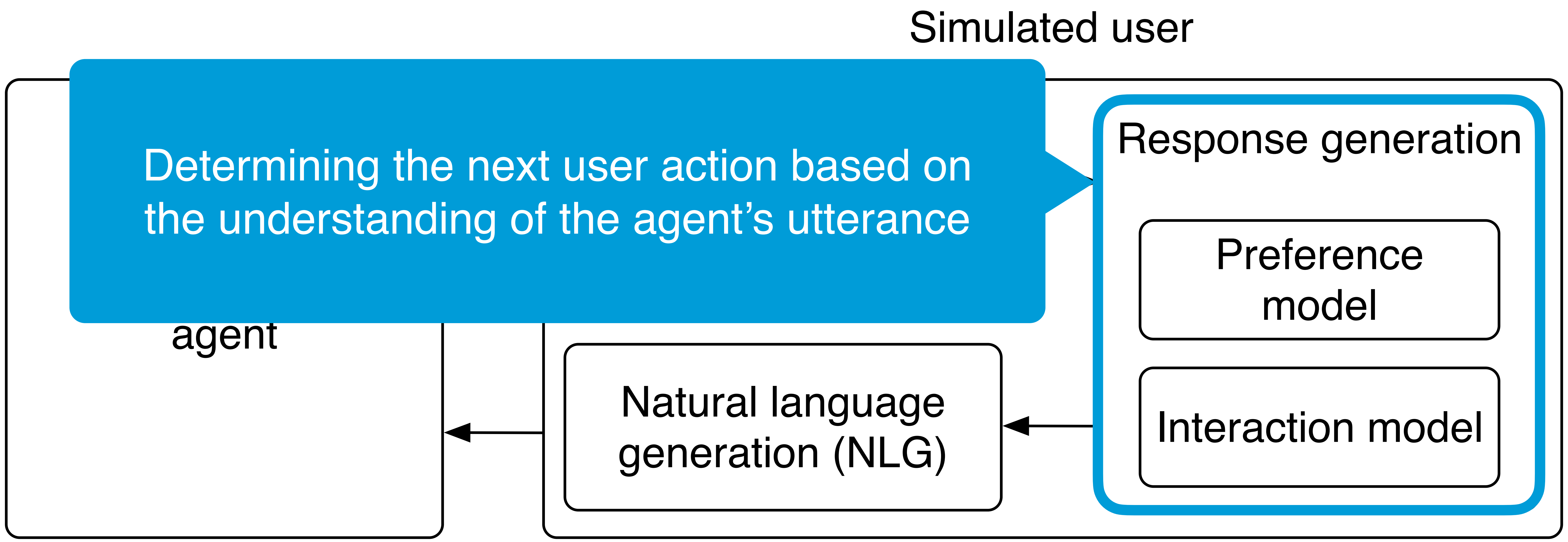
SIMULATION FRAMEWORK



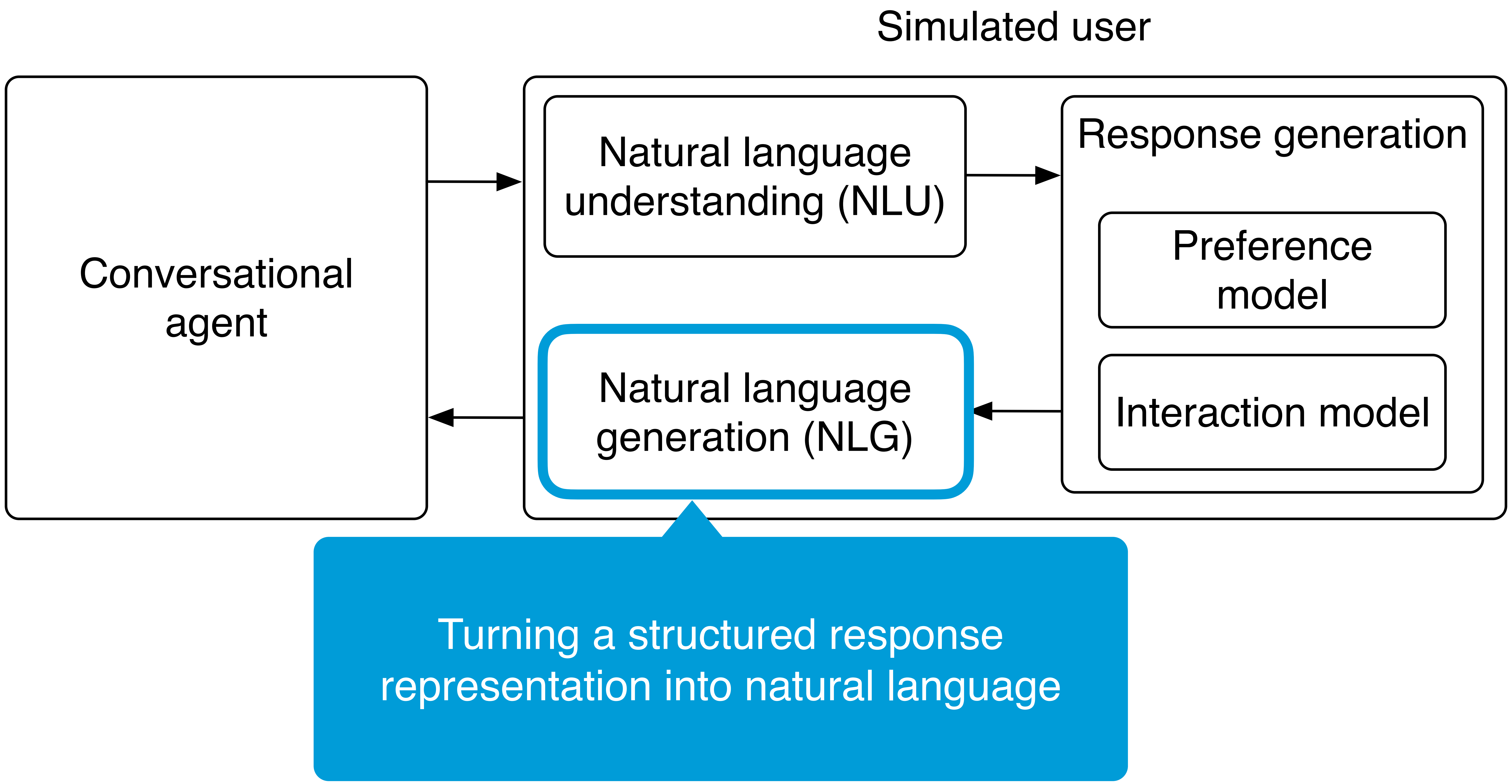
SIMULATION FRAMEWORK



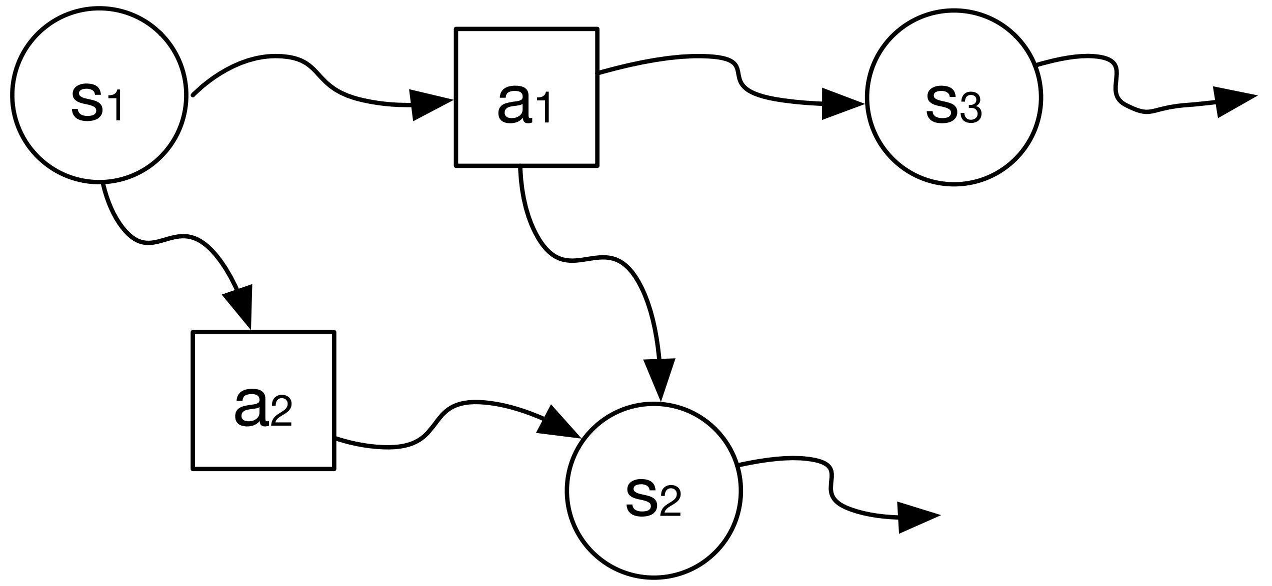
SIMULATION FRAMEWORK



SIMULATION FRAMEWORK



MODELING SIMULATED USERS



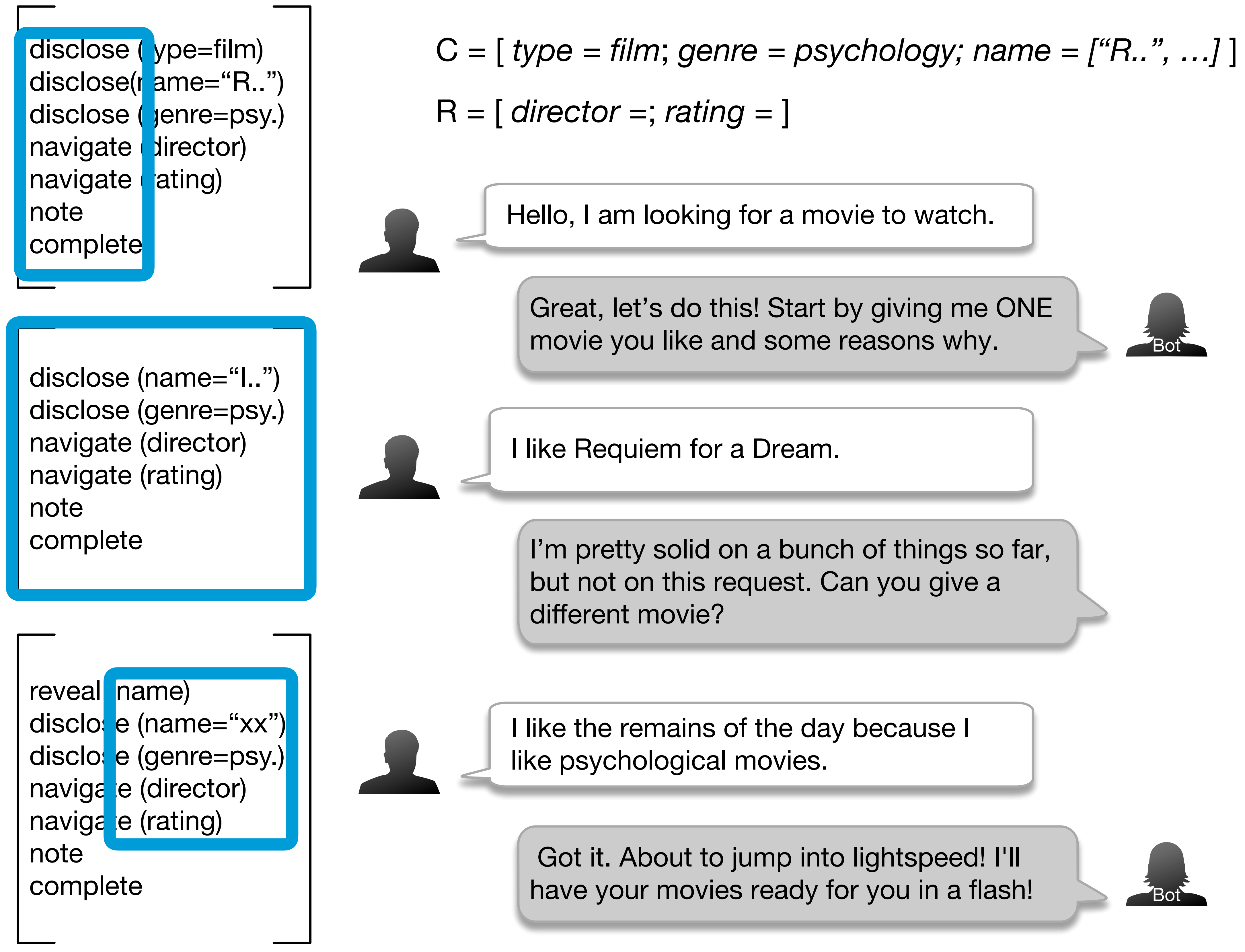
- Model dialogue as a Markov Decision Process
- Every MDP is formally described by a finite

  space S, a finite action set $\mathcal{A}$, and transition probabilities P

  - *Dialogue acts* (or *actions*): task-specific intents that are being communicated in utterances
  - *Dialogue state*: the state of the dialogue manager is in. At each time step (dialogue turn) t, the dialogue manager is in a particular state s_t
  - *Transition probabilities:* the probability of transitioning from s_t to s_{t+1}

AGENDA-BASED SIMULATION*

- The action agenda *A* is a stack-like representation for user actions that is dynamically updated
- The next user action is selected from the top of the agenda
- Agenda updates are regarded as a sequence of pull or push operations
  - Accomplished goal -> pull operation
  - Not accomplished -> push operation



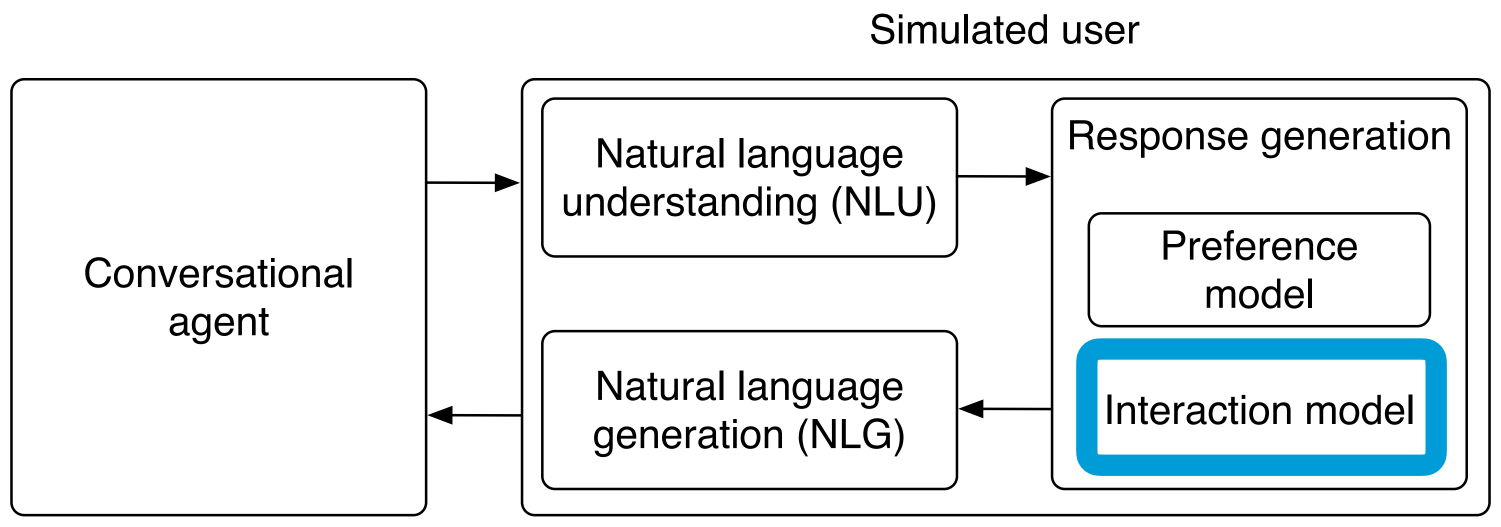
* Schatzmann et al. Agenda-based User Simulation for Bootstrapping a POMDP Dialogue System, NAACL, 2007

ACTION SPACE*

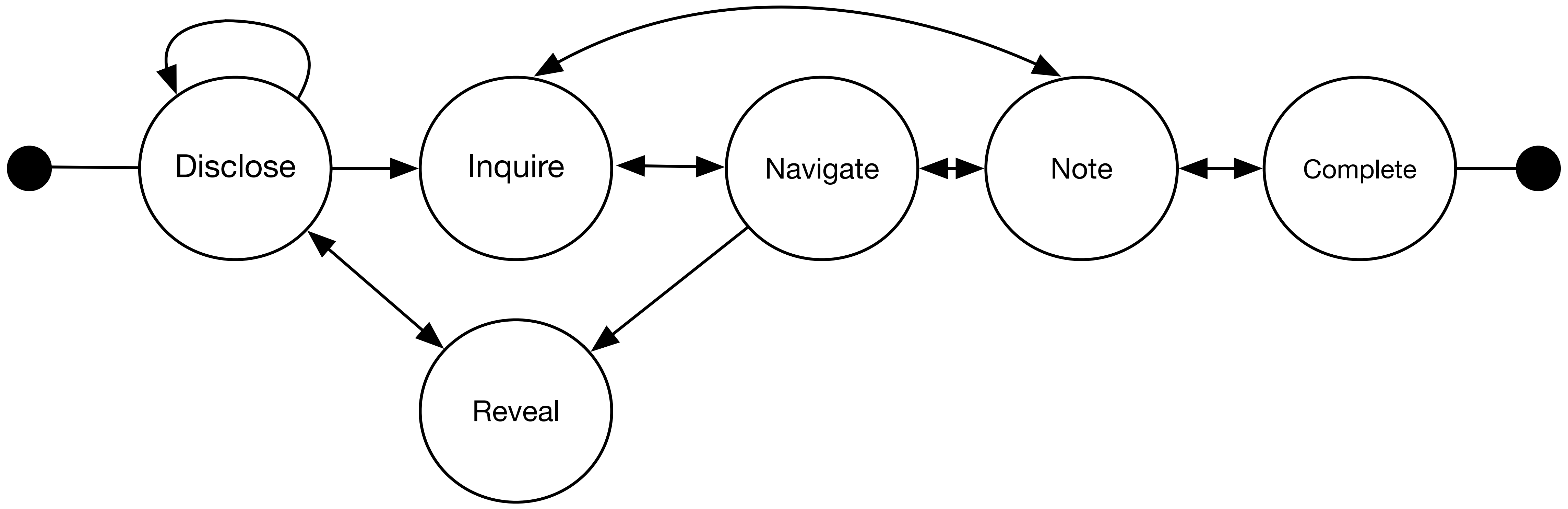
Disclose	I would to arrange a holiday in Italy
Reveal	Actually, we need to go on the 3rd of May in the evening
Inquire	What other regions in Europe are like that?
Navigate	Which one is the cheapest option?
Note	That hotel could be a possibility
Complete	Thanks for the help, bye

* Azzopardi et al. Conceptualizing Agent-human interactions during the conversational search process. CAIR 2018

INTERACTION MODEL

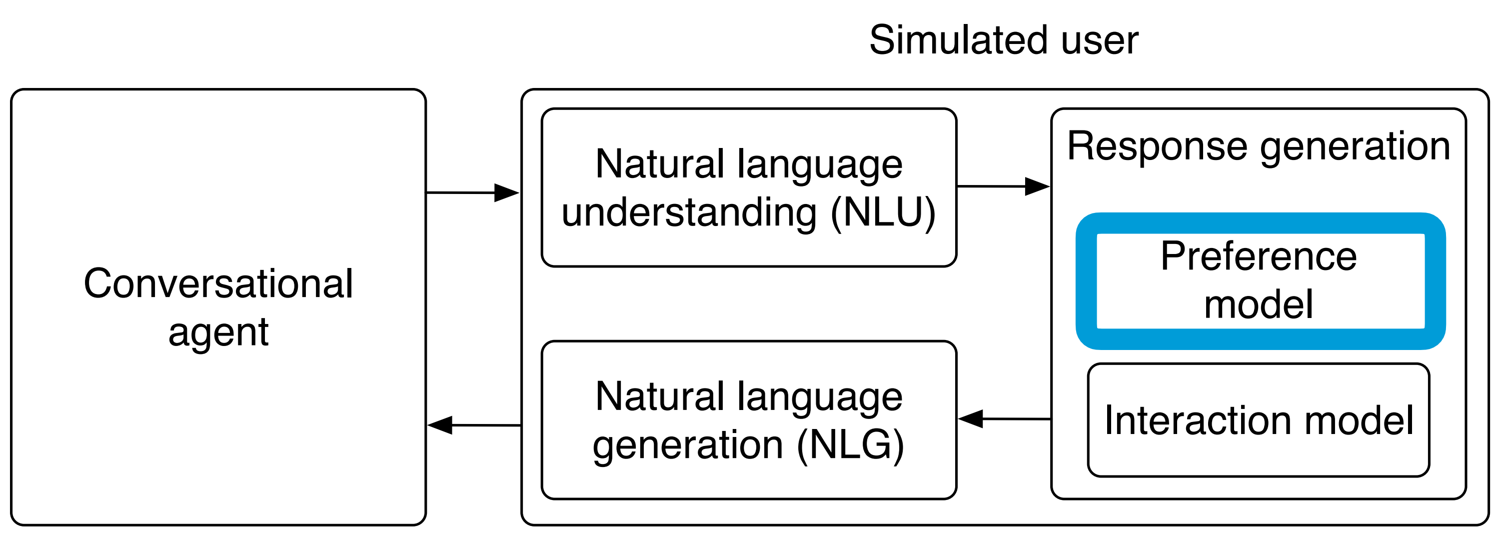


- The interaction model defines how the agenda should be initialized (A_0) and updated ($A_t \Rightarrow A_{t+1}$)
  - QRFA Model *
    - User: Query and Feedback
    - Agent: Request and Answer
    - QRFA are mapped to action space manually
  - CIR6 Model



* Vakulenko et al. QRFA: A Data-Driven Model for Information Seeking Dialogues. ECIR 2019.

PREFERENCE MODEL



- The preference model is meant to capture individual differences and personal tastes
- Preferences are represented as a set of attribute-value pairs
  - Single Item Preference
    - Check if *i* in I_u an answer accordingly, and randomly decide preference
    - It offers limited consistency
  - Personal Knowledge Graph*
    - PKG has two types of nodes: items and attributes
    - Infers the rating for that attribute by considering the ratings of items that have that attribute

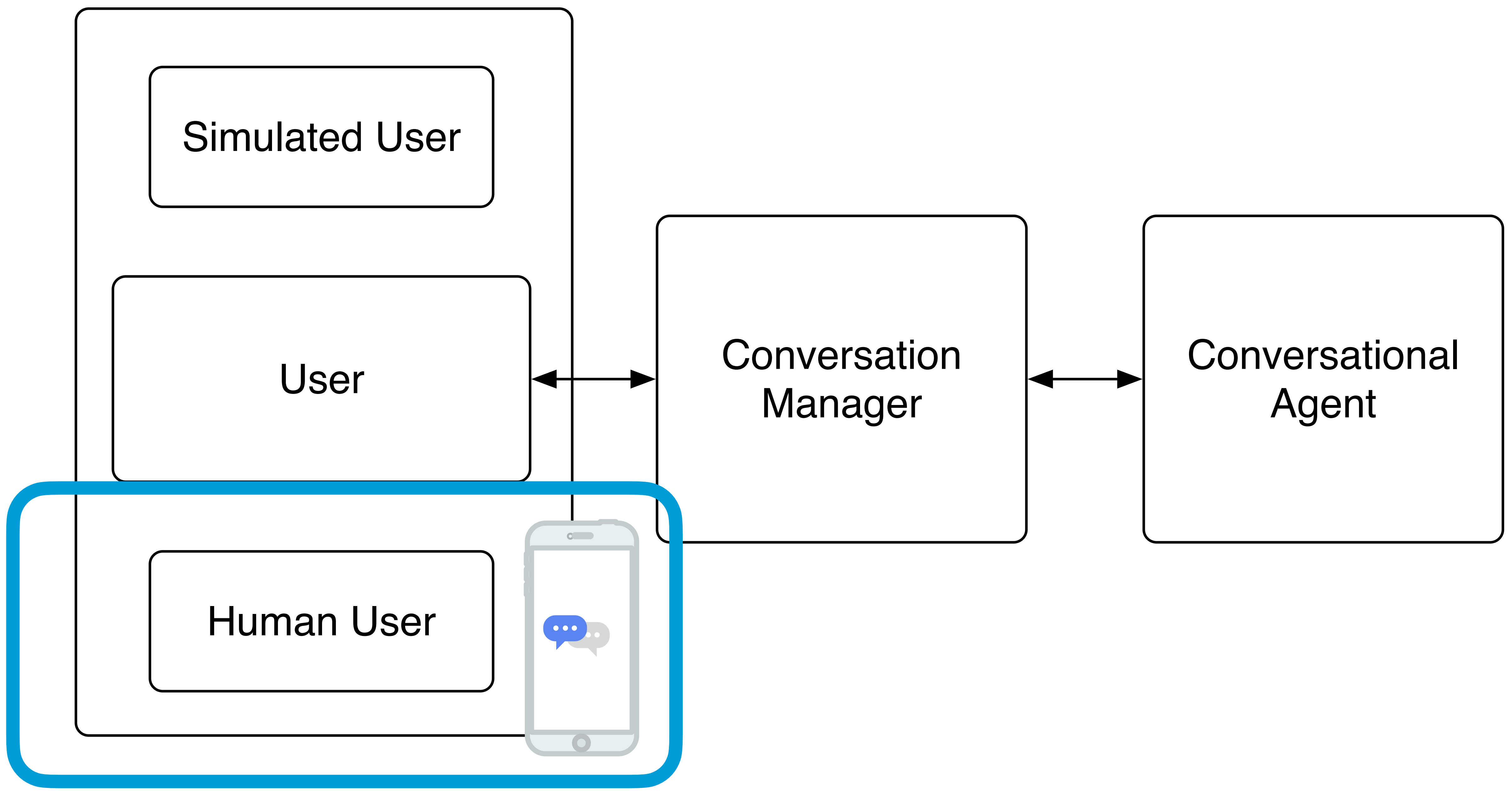
$$r_j = \frac{1}{|I_j|} \sum_{i \in I_j} r_i$$

* Balog et al. Personal Knowledge Graphs: A Research Agenda. ICTIR 2019.

EXPERIMENTAL EVALUATION

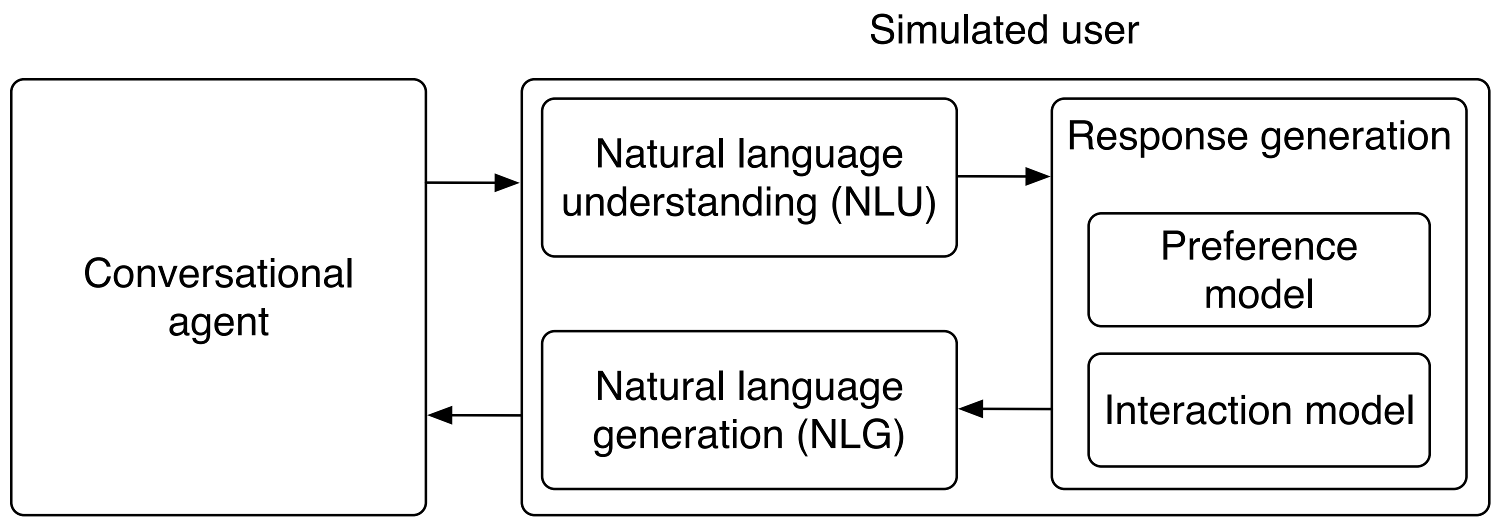
EVALUATION ARCHITECTURE

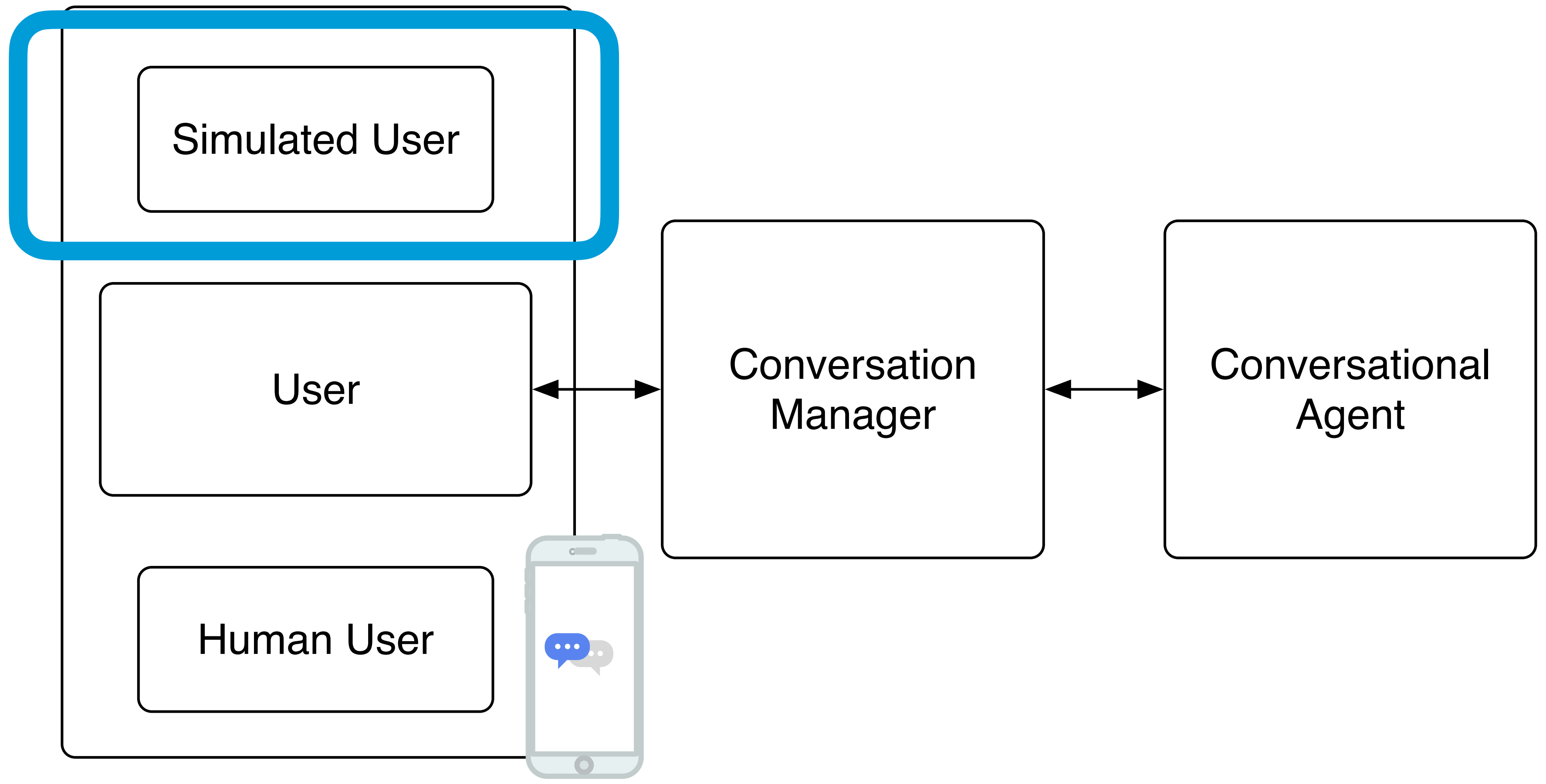
- Three existing conversational movie recommenders (A, B, C) are compared using both real ( ) and simulated ( ) users
  - Real users: we invite crowdsourcing workers to interact with recommenders on Telegram, and use their dialogue records for initializing the simulated users



EVALUATION ARCHITECTURE

- Simulated users
  - Preference model is initialized by sampling historical preferences of a real user from MovieLens data
  - Interaction model is trained based on behaviors of real human users
  - Both NLU and NLG use hand-crafted templates





CHARACTERISTICS OF CONVERSATIONS

- (RQ1) How well do our simulation techniques capture the characteristics of conversations?

Method	AvgTurns			UserActRatio			DS-KL		
	A	B	C	A	B	C	A	B	C
Real users	9.20	14.84	20.24	0.374	0.501	0.500	—	—	—
QRFA-Single	10.52	12.28	17.51	0.359	0.500	0.500	0.027	0.056	0.029
CIR6-Single	9.44	12.75	15.92	0.382	0.500	0.500	0.055	0.040	0.025
CIR6-PKG	6.16	9.87	10.56	0.371	0.500	0.500	0.075	0.056	0.095

- CIR6-PKG tends to have significantly shorter average conversation length, since it terminates the dialog as soon as the user finds a recommendation they like

PERFORMANCE PREDICTION

- (RQ2) How well do the relative ordering of systems according to some measure correlate when using real vs. simulated users?

Method	Reward	Success rate
Real users	A (8.88) > B (7.56) > C (6.04)	B (0.864) > A (0.833) > C (0.727)
QRFA-Single	A (8.04) > B (7.41) > C (6.30)	B (0.836) > A (0.774) > C (0.718)
CIR6-Single	A (8.64) > B (8.28) > C (6.01)	B (0.822) > A (0.807) > C (0.712)
CIR6-PKG	A (11.12) > B (10.65) > C (9.31)	A (0.870) > B (0.847) > C (0.784)

Performance of conversational agents using real vs. simulated users, in terms of Reward and Success Rate. We show the relative ordering of agents (A–C), with evaluation scores in parentheses.

- High correlation between automatic and human evaluations

REALISTICITY

- (RQ3) Do more sophisticated simulation approaches (i.e., more advanced interaction and preference modeling) lead to more realistic simulation?

Method	A			B			C			All		
	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie
QRFA-Single	20	39	16	22	33	20	19	43	13	27%	51%	22%
CIR6-Single	27	30	18	23	33	19	26	40	9	33%	46%	21%
CIR6-PKG	22	39	14	27	29	19	32	25	18	36%	41%	23%

- Our interaction model (CIR6) and personal knowledge graphs for preference modeling both bring improvements

FURTHER ANALYSIS

- We analyze the reasons when the crowd workers chose the real users, and classify them as follows

Style	Realisticity	how realistic or human-sounding a dialog is
	Engagement	involvement of the user in the conversation
	Emotion	expressions of feelings or emotions
Content	Response	user does not seem to understand the agent correctly
	Grammar	language usage, including spelling and punctuation
	Length	the length of reply

SUMMARY OF CONTRIBUTIONS

- A general framework for evaluating conversational recommender agents via simulation
- Interaction and preference models to better control the conversation flow and to ensure the consistency of responses given by the simulated user
- Experimental comparison of three conversational movie recommender agents, using both real and simulated users
- Analysis of comments collected from human evaluation, and identification of areas for future development

THANK YOU!

- Resources: https://github.com/iai-group/UserSimConvRec