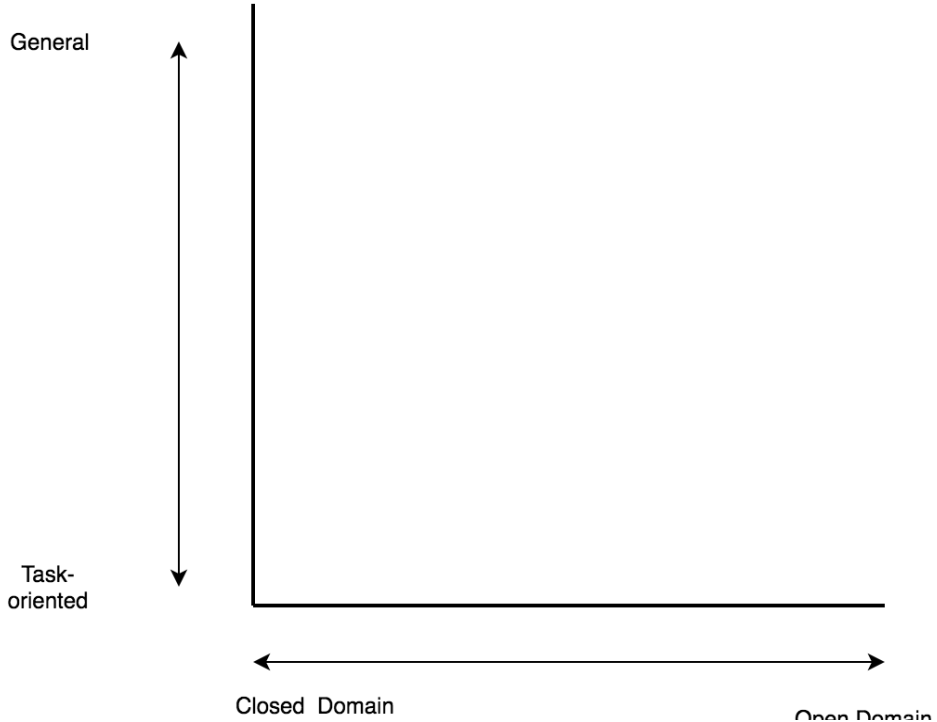
Dialog Systems and How to Score Them

Valentin Malykh, MIPT

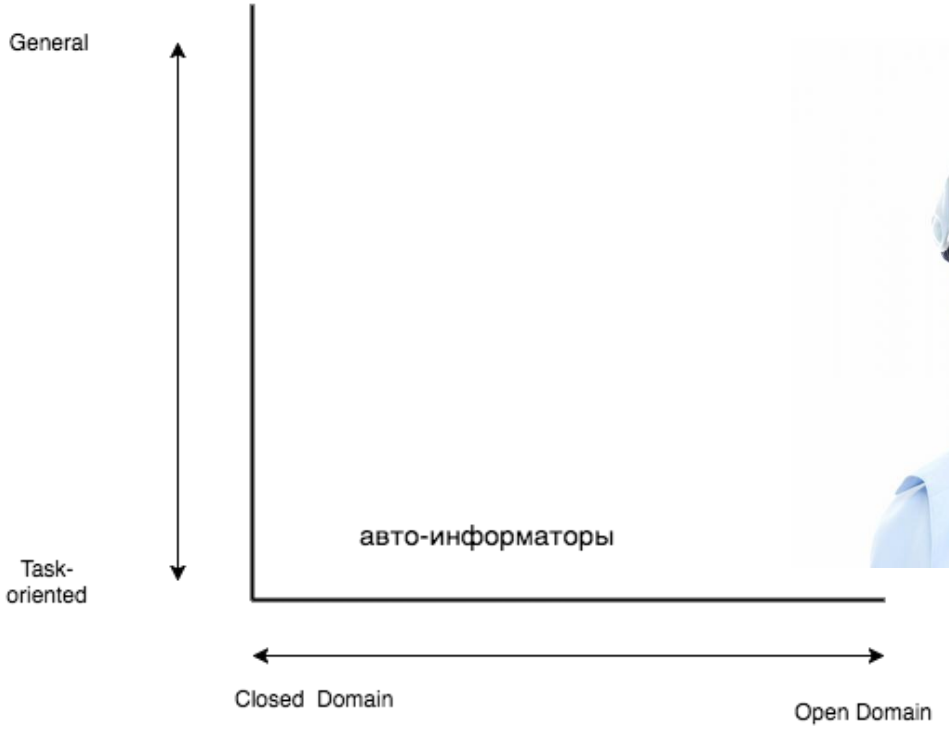
Dialog Systems: Where are they?

Dialog Systems: Examples



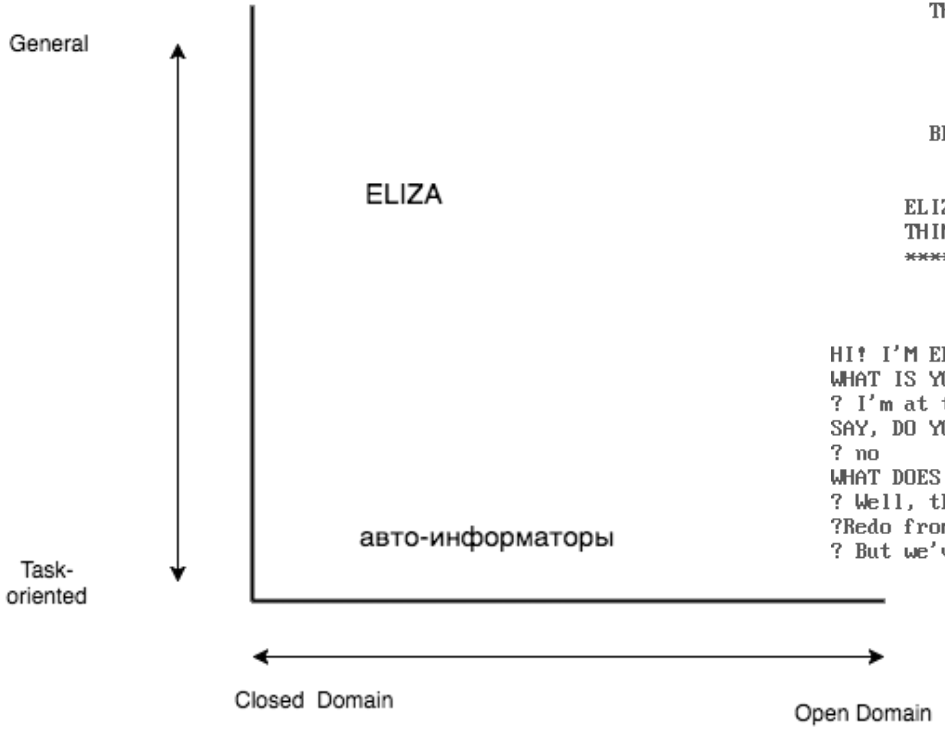
Dialog Systems: Examples

Auto-informer
- Closed Domain
- Task-oriented



Dialog Systems: Examples

ELIZA
- Closed Domain
- General

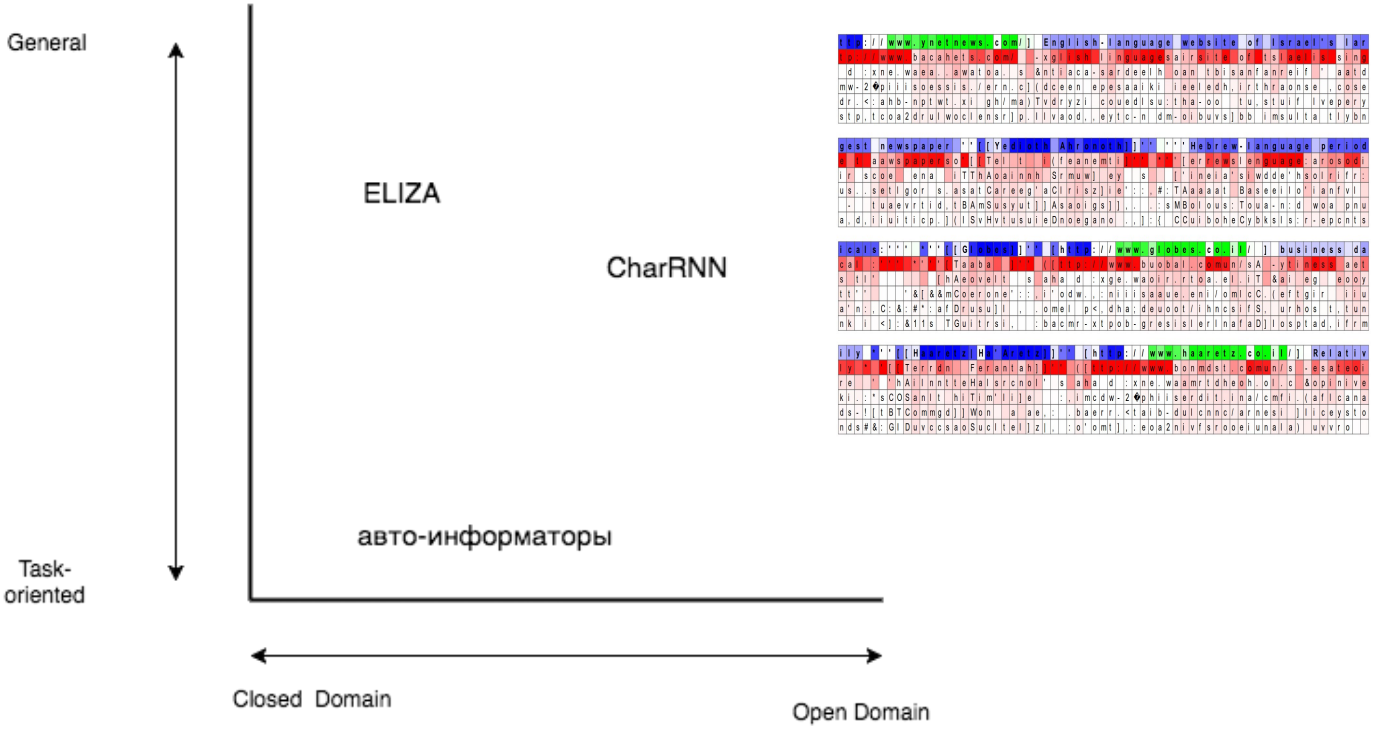



```
**********************************
  THE BROADCASTERS DATABASE
       PRESENTS:

            ELIZA

    BE SURE CAPS LOCK IS ON
      TYPE ! TO TERMINATE
        NO COMMAS PLEASE
  ELIZA TAKES A FEW SECONDS TO
  THINK BEFORE ANSWERING YOU.
  **********************************
```

```
HI! I'M ELIZA THE BROADCAST PSYCHIATRIST,
WHAT IS YOUR PROBLEM?
? I'm at the end of my rope.
SAY, DO YOU WORK IN BROADCASTING?
? no
WHAT DOES THAT SUGGEST TO YOU?
? Well, that I work in some other field.
?Redo from start
? But we've come so far!_
```

Dialog Systems: Examples

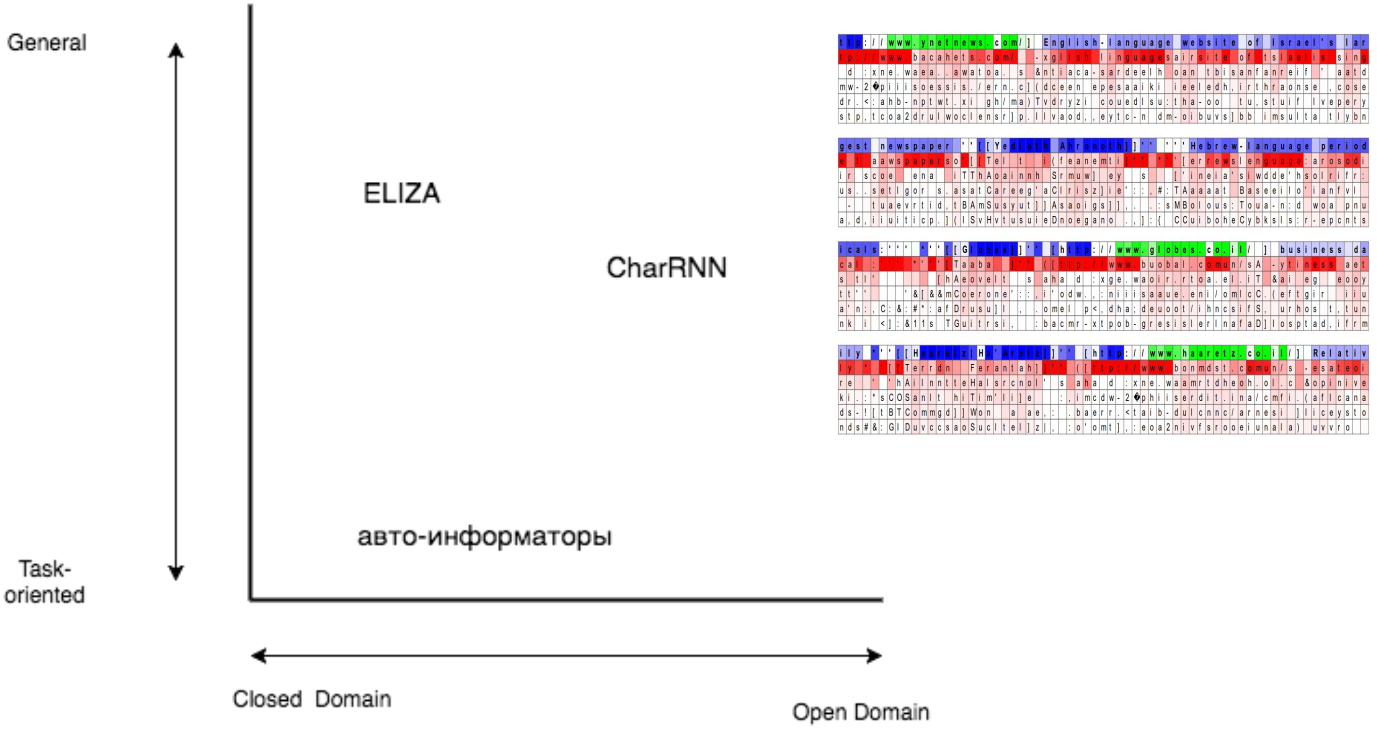
Karpathy's CharRNN
- Open-Domain
- General



Dialog Systems: Examples

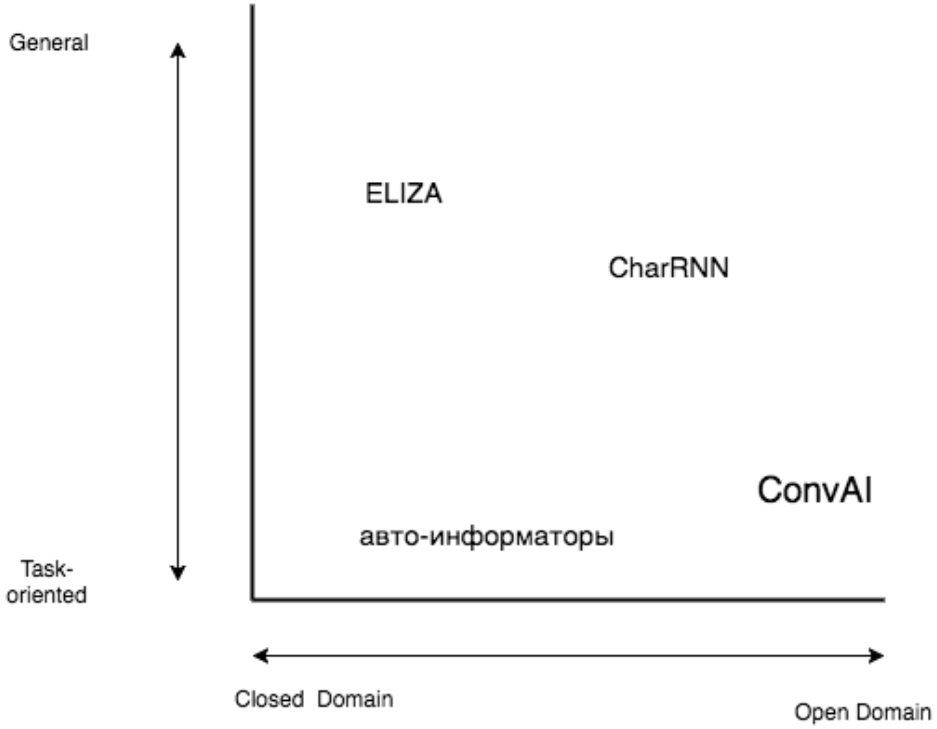
Karpathy's CharRNN
- Open-Domain
- General

- Reactive
- Not that "smart"



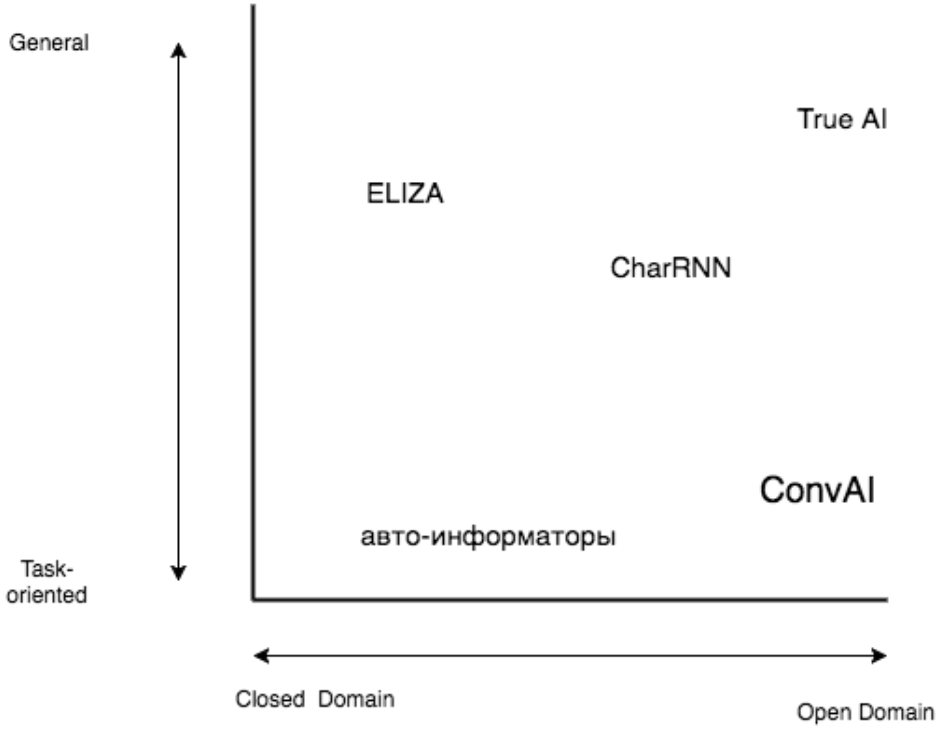
Dialog Systems: Examples

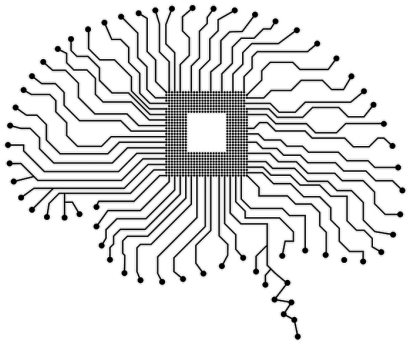
ConvAI Goal
- Open Domain
- Task-oriented



Dialog Systems: Examples

True AI
- Open Domain
- General

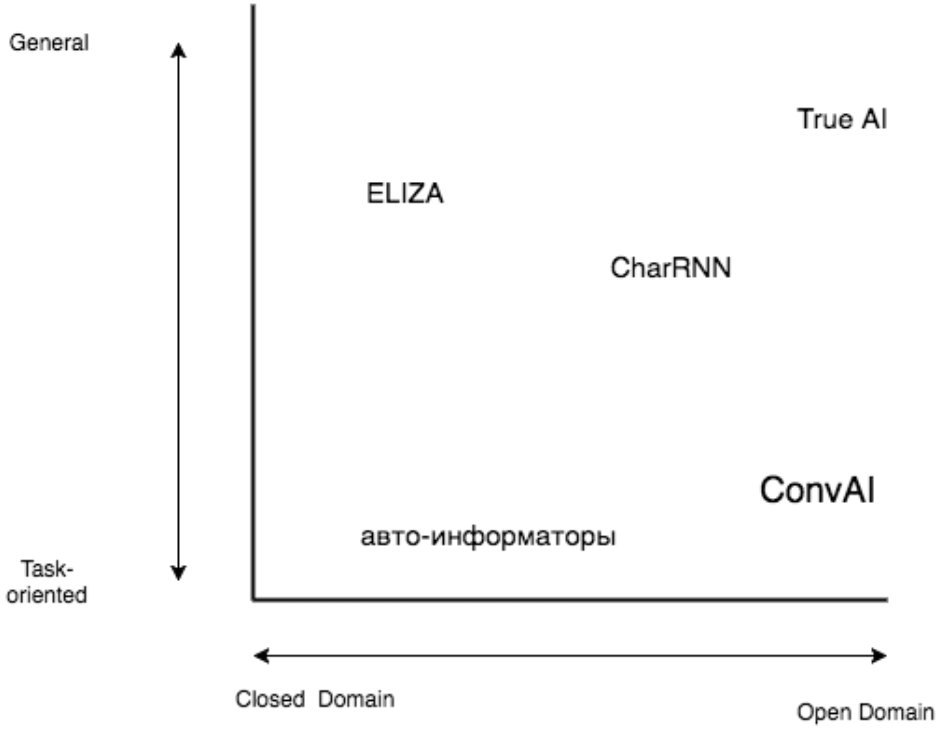


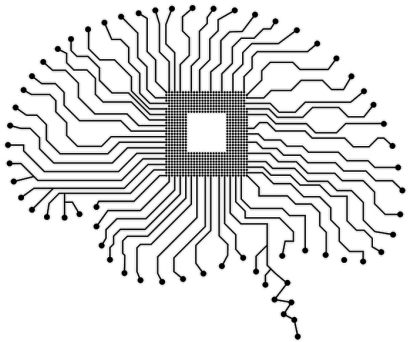


Dialog Systems: Examples

True AI
- Open-Domain
- General

- Pro-active





DIalog Systems: How to Score Them?

Task-oriented Systems

Task Completion Rate (TCR)

$$\text{Task Completion Rate} = \frac{\text{\# Successful Runs}}{\text{\# All Runs}}$$

End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning

Jason D. Williams, Geoffrey Zweig, *arxiv:1606.01269*

General Systems

- Word-Overlap
  - ROUGE
  - BLEU
  - METEOR

$$P_n(r, \hat{r}) = \frac{\sum_k \min(h(k, r), h(k, \hat{r}_i))}{\sum_k h(k, r_i)}$$

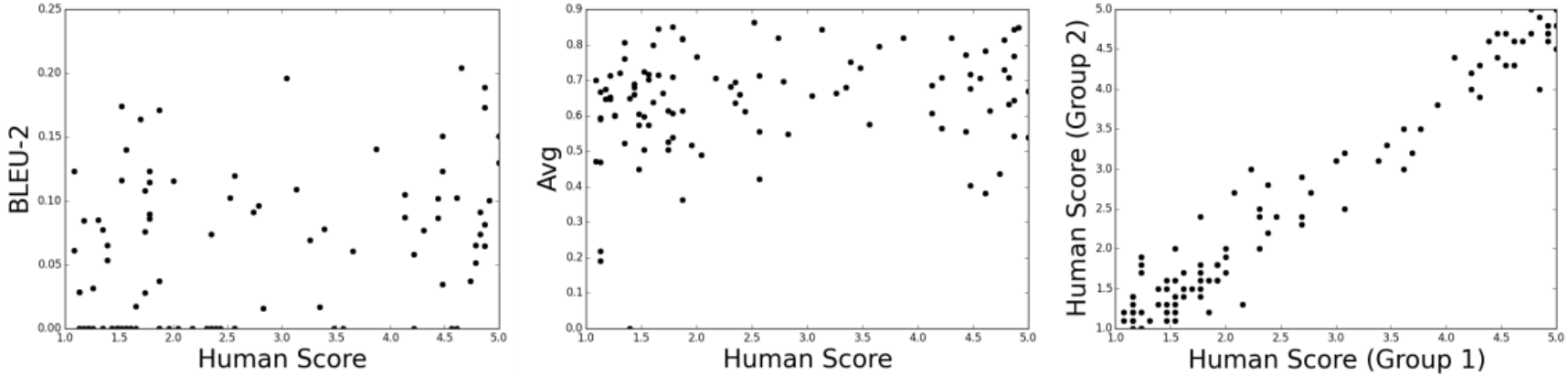
$$\text{BLEU-N} := b(r, \hat{r}) \exp(\sum_{n=1}^{N} \beta_n \log P_n(r, \hat{r}))$$

General Systems

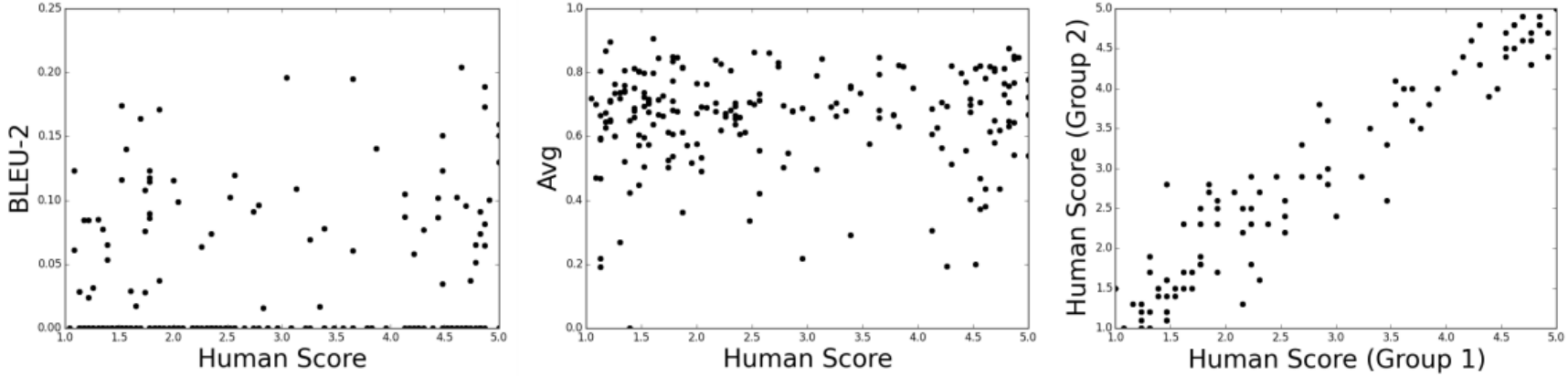
- Word-Overlap
  - ROUGE
  - BLEU
  - METEOR
- Embedding-based
  - Greedy Matching
  - Embedding Average
  - Extreme Vector

$$\bar{e}_r = \frac{\sum_{w \in r} e_w}{\left| \sum_{w' \in r} e_{w'} \right|},$$

$$\mathrm{EA} := \cos(\bar{e}_r, \bar{e}_{\hat{r}})$$



(a) Twitter



(b) Ubuntu

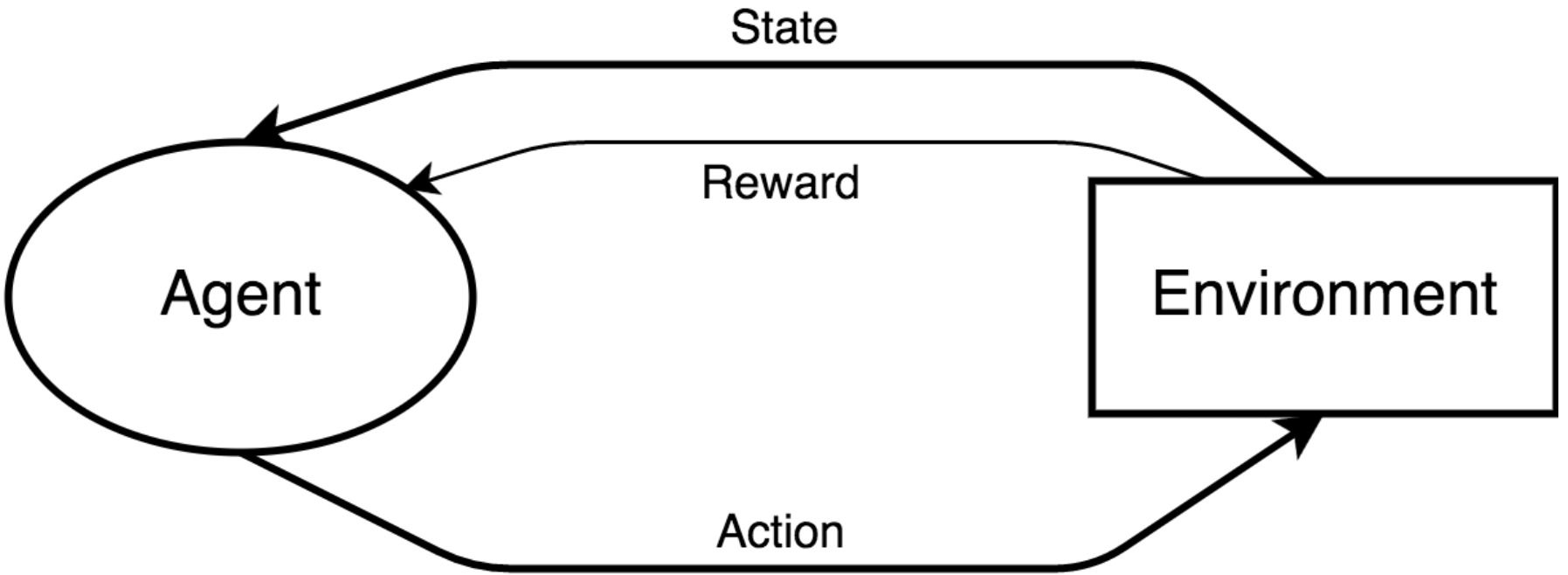
General Systems

How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, Joelle Pineau, *arxiv:1603.08023*

How to Live with That?

How to Live with That: Reinforcement Learning



How to Live with That

- Reward-based Imitation (RBI)
- Forward Prediction (FP)

Dialog-based Language Learning

Jason Weston, *arXiv:1604.06045*

How to Live with That

Model	$r = 0$	$r = 0.1$	$r = 0.5$	$r = 1$
Reward Based Imitation (RBI)	0.333	0.340	0.365	0.375
Forward Prediction (FP)	0.358	0.358	0.358	0.358
RBI+FP	0.431	0.438	0.443	0.441

Dialogue Learning With Human-In-The-Loop

Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, Jason Weston, *arxiv:1611.09823*

How to Live with That

bAbI Task 6: Partial Rewards

Mary went to the hallway.
John moved to the bathroom.
Mary travelled to the kitchen.
Where is Mary? kitchen
Yes, that's right!
Where is John? bathroom
Yes, that's correct! (+)

WikiMovies Task 6: Partial Rewards

What films are about Hawaii? 50 First Dates
Correct!
Who acted in Licence to Kill? Billy Madison
No, the answer is Timothy Dalton.
What genre is Saratoga Trunk in? Drama
Yes! (+)
...

Conversational Intelligence Challenge

- Task-oriented
- Open Domain

NIPS 2017 Live Competition

Open Source Dataset as a Result

Presumed Dataset Size: 1000+ dialogs

convai.io

Conversational Intelligence Challenge

Organizers:

Mikhail Burtsev, Valentin Malykh, *MIPT, Moscow*

Ryan Lowe, *McGill University, Montreal*

Iulian Serban, Yoshua Bengio, *University of Montreal, Montreal*

Alexander Rudnicky, Alan W. Black, Shrimai Prabhumoye, *Carnegie Mellon University, Pittsburgh*

convai.io

Conversational Intelligence Challenge

- Task-oriented
- Open Domain

Person chats with a bot (or another person)

convai.io

Conversational Intelligence Challenge

- Task-oriented
- Open Domain

Person chats with a bot (or another person)

They are discussing a news article

convai.io

Conversational Intelligence Challenge

- Task-oriented
- Open Domain

Person chats with a bot (or another person)

They are discussing a news article

Human judgment at the end (consistency, overall adequacy)

convai.io

Conversational Intelligence Challenge

The Judgement Scheme

- Like/dislike for each line of a bot
- End scoring:
  - quality
  - breadth
  - engagement

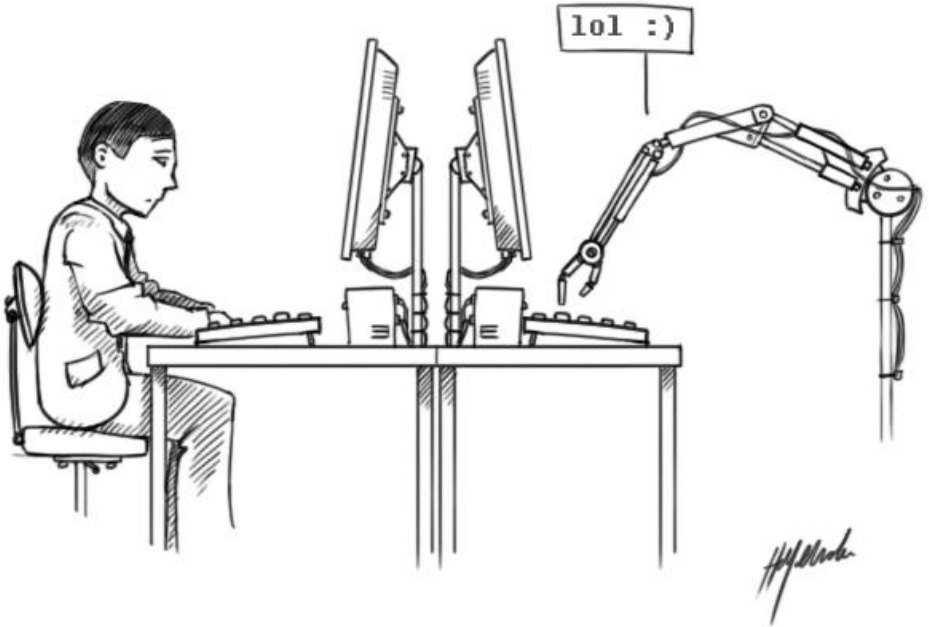
convai.io

Turing Test

Two persons are talking indirectly

⅔ of persons who talked to a bot

should say that it’s a human



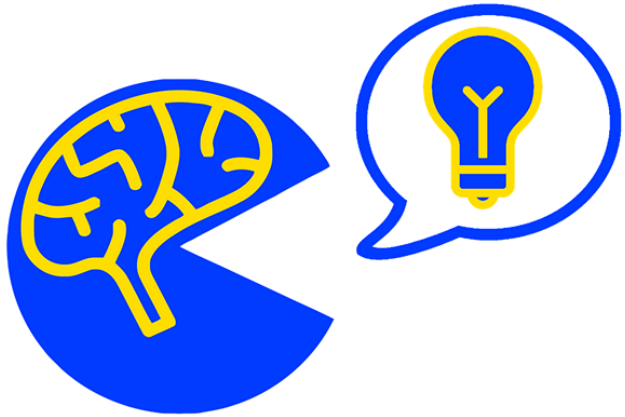
Summer School & Hackathon

The task is to determine if the dialog participant is bot or a human

Speakers from leading academia & industry labs: last time there were DeepMind, Facebook, ETHZ

Last week of July

Get the updates at deephack.me!



iPavlov Project

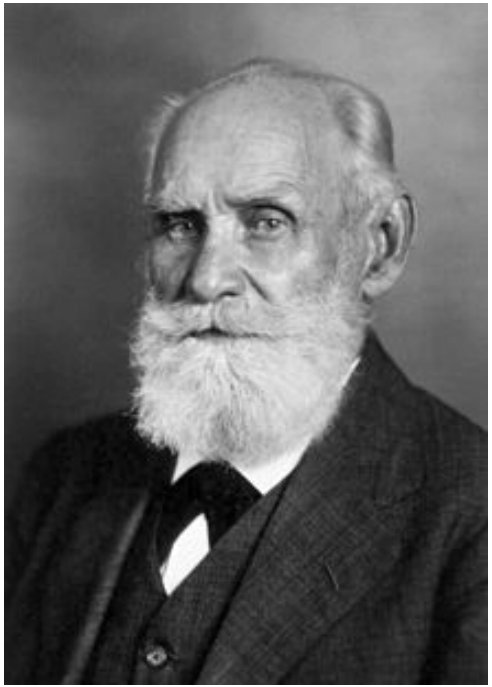
Devoted to develop an open source library with building blocks for (almost) any NLP task

Funded by National Technology Initiative and Sberbank

3 years

Open to collaboration

iPavlov.ai



Summary

For task-oriented systems TCR is great

For any systems classical approaches are not so great

But we can get use of user models in RL

Also we invite everyone to participate in Summer School & ConvAI.io

Questions?