Course
Knowledge-Based Systems

Lecture 13 - Procedural Knowledge and Wrap up

Dr. Mohsen Mesgar

Universität Duisburg-Essen

Recall: Types of Knowledge

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

What we have seen so far:

- Factual knowledge
- Commonsense knowledge
- Categorial knowledge

→ These are **declarative knowledge** types that you could store in databases (as a machine) or in your declarative memory (as a human)

But how about:

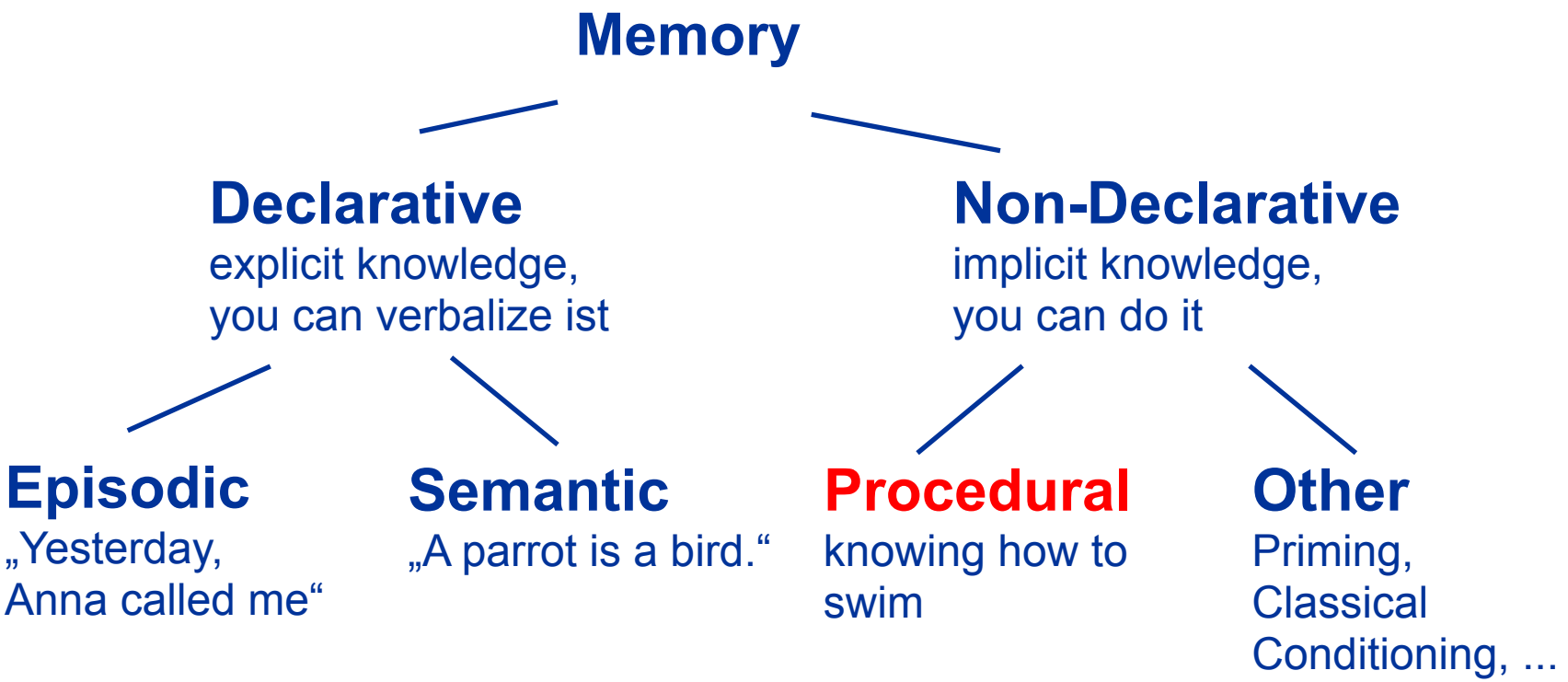
- riding a bike or ice-skating
- flipping a pancake or playing the violin?
- play chess and actually win?

→ **Procedural knowledge!** As a human, you have to learn how to do it by doing it (trial and error). As a machine, you can ideally learn it in the same way.

Squire‘s Memory Classification (for Humans)

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



How to Acquire Procedural Knowledge

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

For humans:

- Acquisition of procedural knowledge is a complex process, involving practice (**trial and error**). Imagine a child learning to walk, falling down, getting up again, fall again...
- Learning procedural knowledge means learning by **interacting with your environment.** („Oh, I fell down something must have gone wrong!“)
- That means we learn by receiving positive or negative feedback **(„Reinforcement“).** Feedback could be for the learning to walk scenario:
  - falling down and hurting yourself
  - still be standing and walking and not get hurt
  - finally reach the table where the chocolate is
- but also:
  - winning or loosing a game of chess
  - finish Vivaldi‘s violin concerto without your neighbors complaining

Reinforcement Learning

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- Learning by interacting with your environment.

https://images.contentful.com/6m9bd13t776q/6ErMnXakpimYOMcUI02YY0/498b44585d4c374aa57c662138f3739f/baby-walking-shoes-desktop.jpg?q=75

A Machine Learning Procedural Knowledge

- The pancake flipping robot

Reinforcement Learning

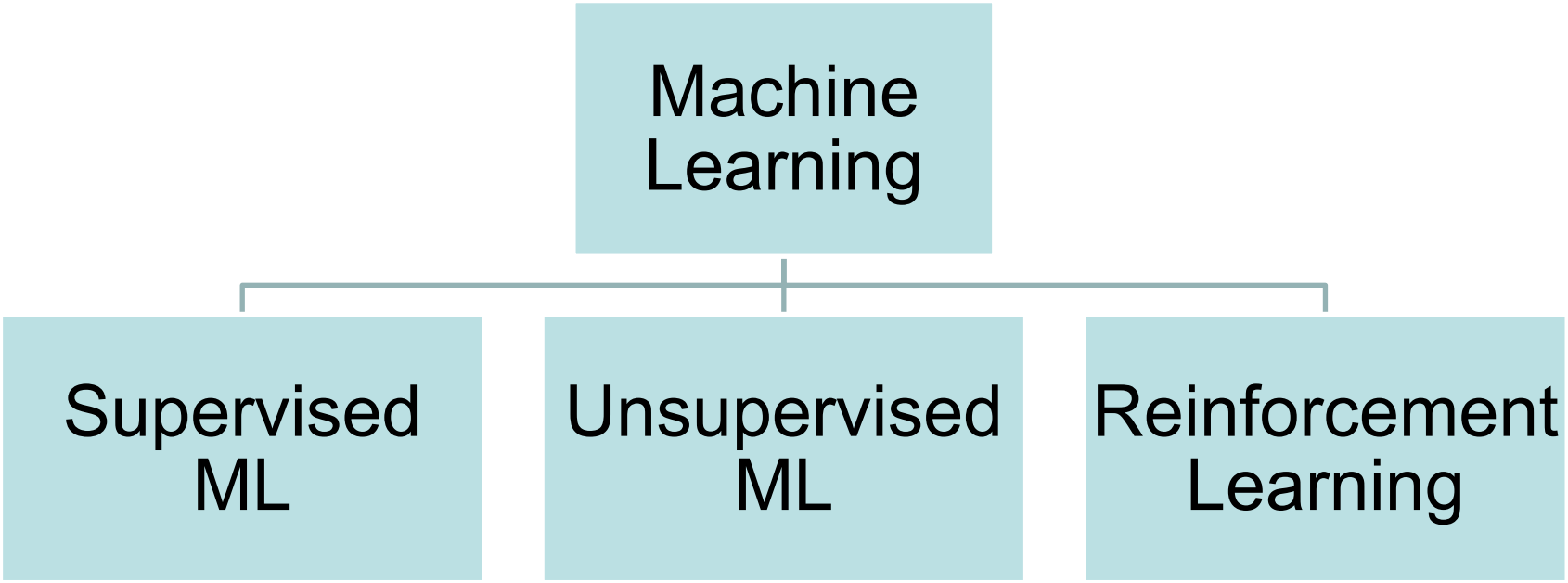
UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

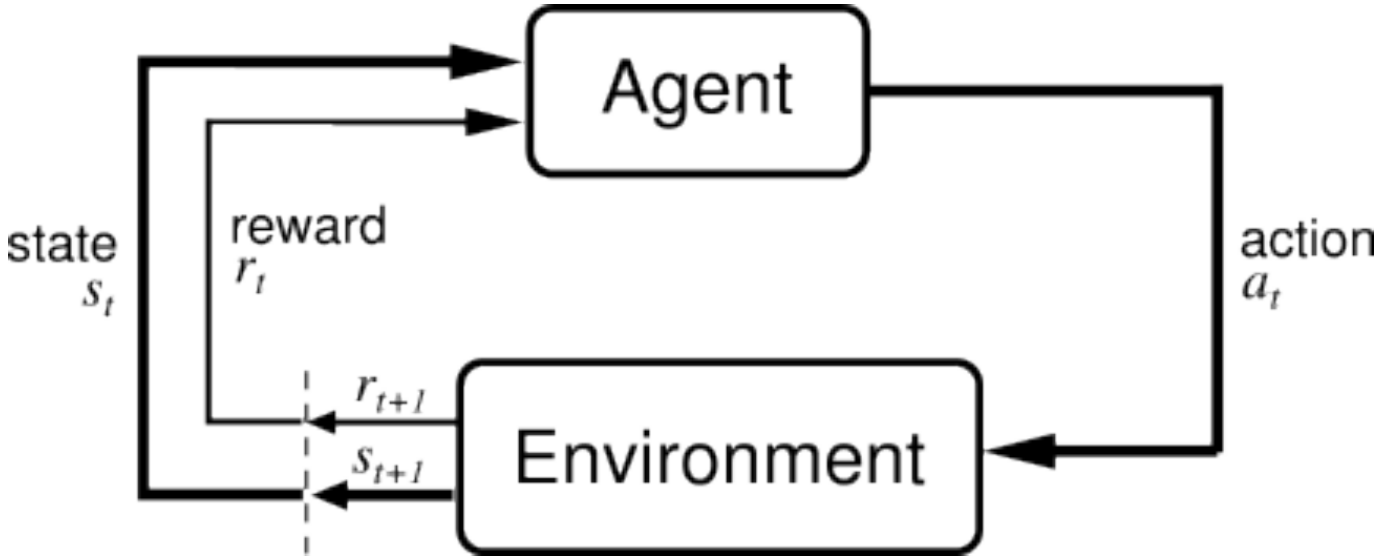
- Learning by interacting with your environment.
  - Receive a reward (or punishment) as an estimate of how close the agent is to a goal
  - Learn the best strategy by maximizing the total average reward

http://3.bp.blogspot.com/_1fjMGlWo2xc/TIpubgHQIKI/AAAAAAAAAO8/vXQiCFtbgbk/s400/mouse+maze.jpg

Types of Machine Learning



Reinforcement Learning (RL): Interaction with the Environment



Sutton & Barto: Reinforcement Learning

In RL, an agent learns by constantly interacting with an environment. The agent in a certain **state** performs an **action**. By performing this action it receives a **reward** and reaches a **new state.** Then it performs the next action. Learning in this scenario means to **choose the next action so that the sequence of actions maximize total received reward**.

Reinforcement Learning: Core Concepts

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Let‘s assume we want to learn our way through the maze to get the cheese:

- **Action a_t**: what the agent does at a certain time step t,
  e.g., move forward or turn left
- **State s_t**: representation of the current environment at timestep t,
  e.g., position in the maze
- **Reward r_t**: numerical, received based on an action in a certain state at time t, e.g., *-0.1 (no cheese)* vs *1 (cheese)*
  (Getting the cheese gets you a big reward, one more step without cheese only a small punishment, but that is a design decision)
- **Policy π**: a strategy telling us which action to take in each possible state:
  - *e.g. for square 1 in the maze:*
    P(turn left) = 0.5
    P(turn right) = 0.25
    P(walk straight) = 0.25

  (*Strategies can often be probabilistic.* There could also be a fix strategy that tells you exactly the only one action you can do in each step)

Overall goal: Find a ***policy*** π which **maximizes the total reward**, e.g. *get the cheese as quickly as possible*

What is a good strategy? Exploration vs. Exploitation

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- You are standing in front of a number of slot machines („one-armed bandits“). Each machine has a fixed probability to hit the jackpot. You have a lot of time to play. For free. What would you do to maximize the money you get?

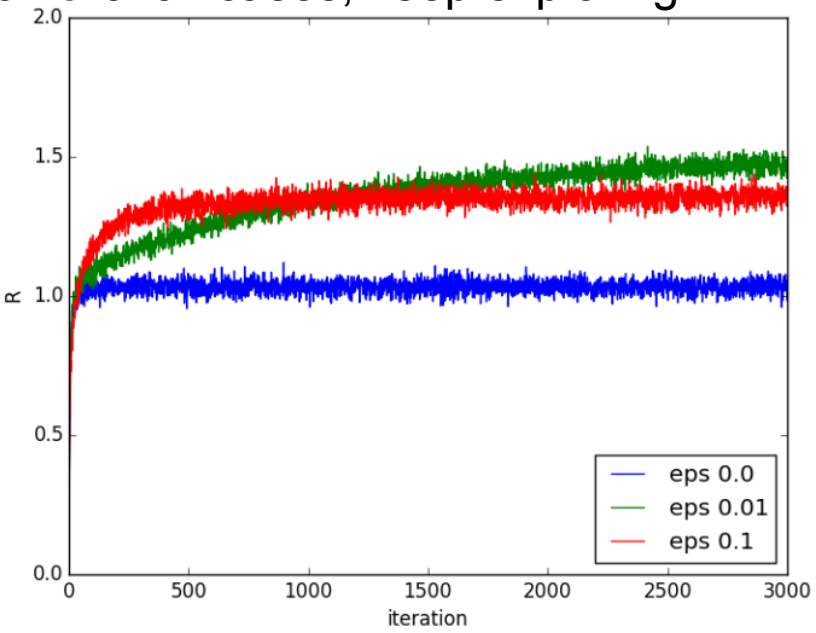
- You keep track of the average outcome for each machine:
  - **exploitation**: select a **greedy action –** choose the machine that promises the highest reward
  - **exploration**: select a non-greedy action, to get a better estimate of the other machines

Trade-off between Exploration and Exploitation

UNIVERSITÄT DUISBURG ESSEN

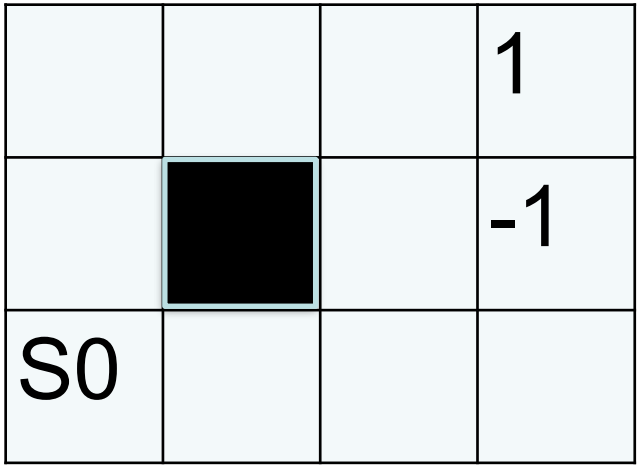
Offen im Denken

ε-greedy strategy: select an action greedily most of the time, but for some fraction ε of all cases, keep exploring



In RL, we often plot reward over time, where one iteration is one action. In the graph you can see, that we get a higher reward earlier if we do more exploration (higher epsilon), but in the long term, selecting the best action greedily is better. A good strategy could be to have a higher epsilon initially (more exploration) and then use a lower epsilon (more exploitation).

The Gridworld Example



Possible Actions with probability 0.8:

right 

 left

down 

 up

Start from S0
Goal state is +1 and -1 is a dead end.

12 **states**

Rewards:
leave the grid: -1
leave state A: +10
leave state B: +5
otherwise: 0

The Gridworld Example

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

0	0	0	1
0		0	-1
0	0	0	0

Evaluate a random policy π:
choose every action with the same probability

Overall expected reward:

$R_t = r_{t+1} + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} + \ldots$

γ: discount parameter, how important is a future reward

(The overall reward is the sum of reward you will get over time. If you play forever, of course you will get an indefinite reward. Therefore, we *discount* future rewards, so that an early reward is more important.)

The Gridworld Example

3.3	8.8	4.4	1
1.5		2.3	-1
-1.9	-1.3	-1.2	-2.0

Evaluate a random policy π: choose every action with the same probability

Overall reward:

$R_t = r_{t+1} + \gamma\, r_{t+2} + \gamma^2\, r_{t+3} + \ldots$

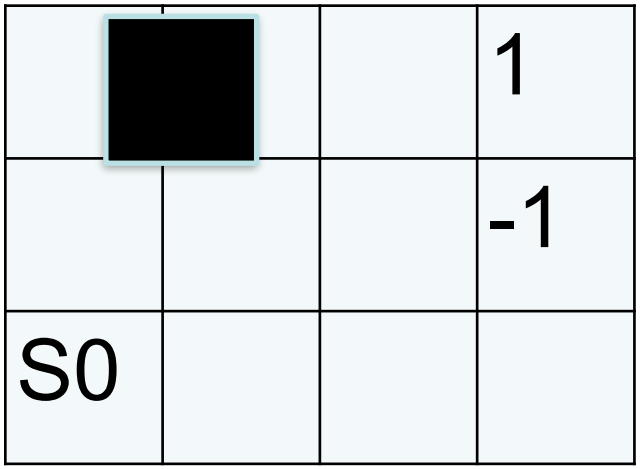
γ: discount parameter, how important is a future reward

Each cell contains the value function $V\pi(s)$ with the overall expected reward

Solving the gridworld problem

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken



To solve the gridworld problem, we want to find the **optimal value function V*** for an **optimal policy π^***

π^* has an **optimal action-value function Q***, telling us the value of every action at a given state

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

We determine Q* iteratively.

Q-learning: Solving the gridworld problem

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Pseudocode:

- Initialize Q(s,a) arbitrarily
- Repeat for each episode:
  - choose a random start state s
  - Repeat for each step of the episode:
    - Choose action a from s using an ϵ greedy policy given the current Q
    - Take action a, observe reward r and next state s‘
    - Update Q based on observed reward r and future reward predicted for s‘
    - Update s to s‘

$$Q(s,a) = (1-\alpha)Q(s,a) + a[r + \gamma \max_{a'} Q(s',\ a')]$$

α: learning rate

$\max_{a'} Q(s',\ a')$:future reward if we take the best action in the next state

Recent Examples of Reinforcement Learning

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- October 2015, AlphaGo defeated the European Go champion Fan Hui.
- Large databse with moves + reinforcement learning by playing against another version of itself.

Recent Examples of Reinforcement Learning

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- **Robotics:** robots learn how to walk
- **Health care:** Find optimal treatment policies for patients with chronic diseases:
- **Dialogue systems**: Find a good dialogue policy for a chatbot.
- **Intelligent tutoring systems**: adapt teaching strategy to optimize learning gain
- **Job shop scheduling:** assign tasks to machines

Summary Reinforcement Learning

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

- RL mimics the way **humans acquire procedural knowledge** through trial and error
- As a variant in machine learning, it is suitable for **problems where it is hard to provide training instances with labels**. Instead we use the reward, that sometimes comes with a delay. (When playing chess we might not know how good an individual move is, but we know in the end whether we won the game.)
- The knowledge of a trained RL system is the **strategy** that tells us what the agent should do in a certain state.
- When modeling a task as RL problem as a developer, we have to carefully decide how to **model the environment**
  - what are allowed action?
  - what states can the agent be in?
  - what reward do we receive for which action in which state?

RL for giving personality to conversational agents

Turing Test

wikipedia.com

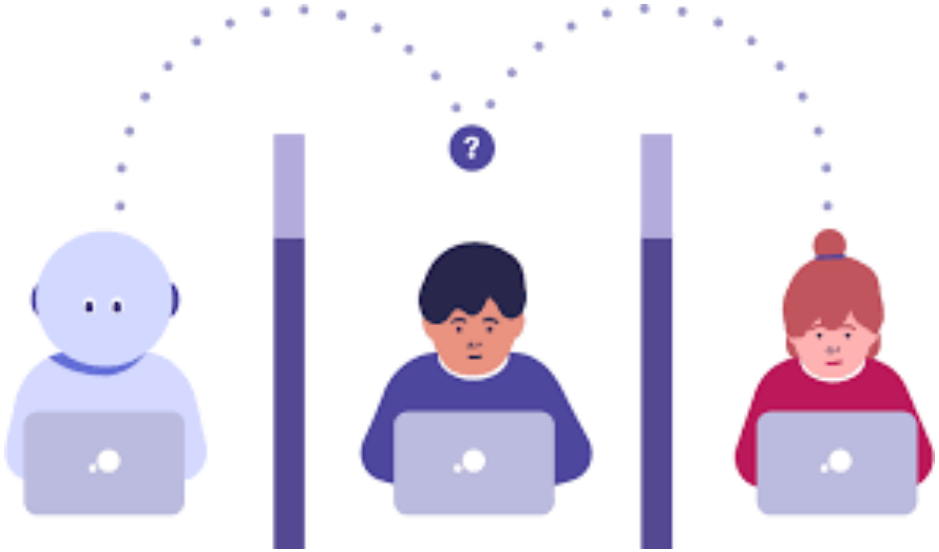
Turing Test

wikipedia.com

Turing Test

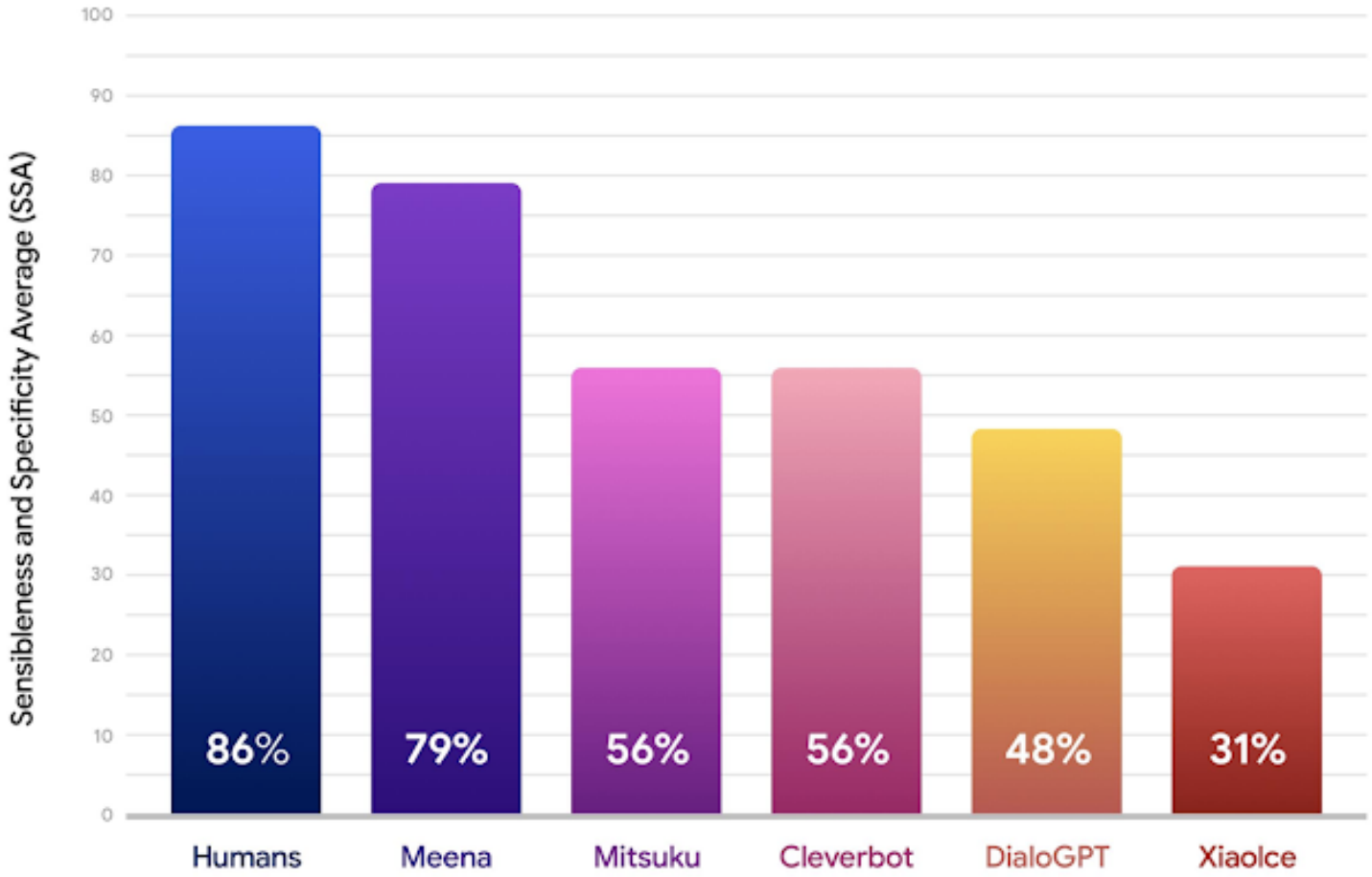
wikipedia.com

Turing Test



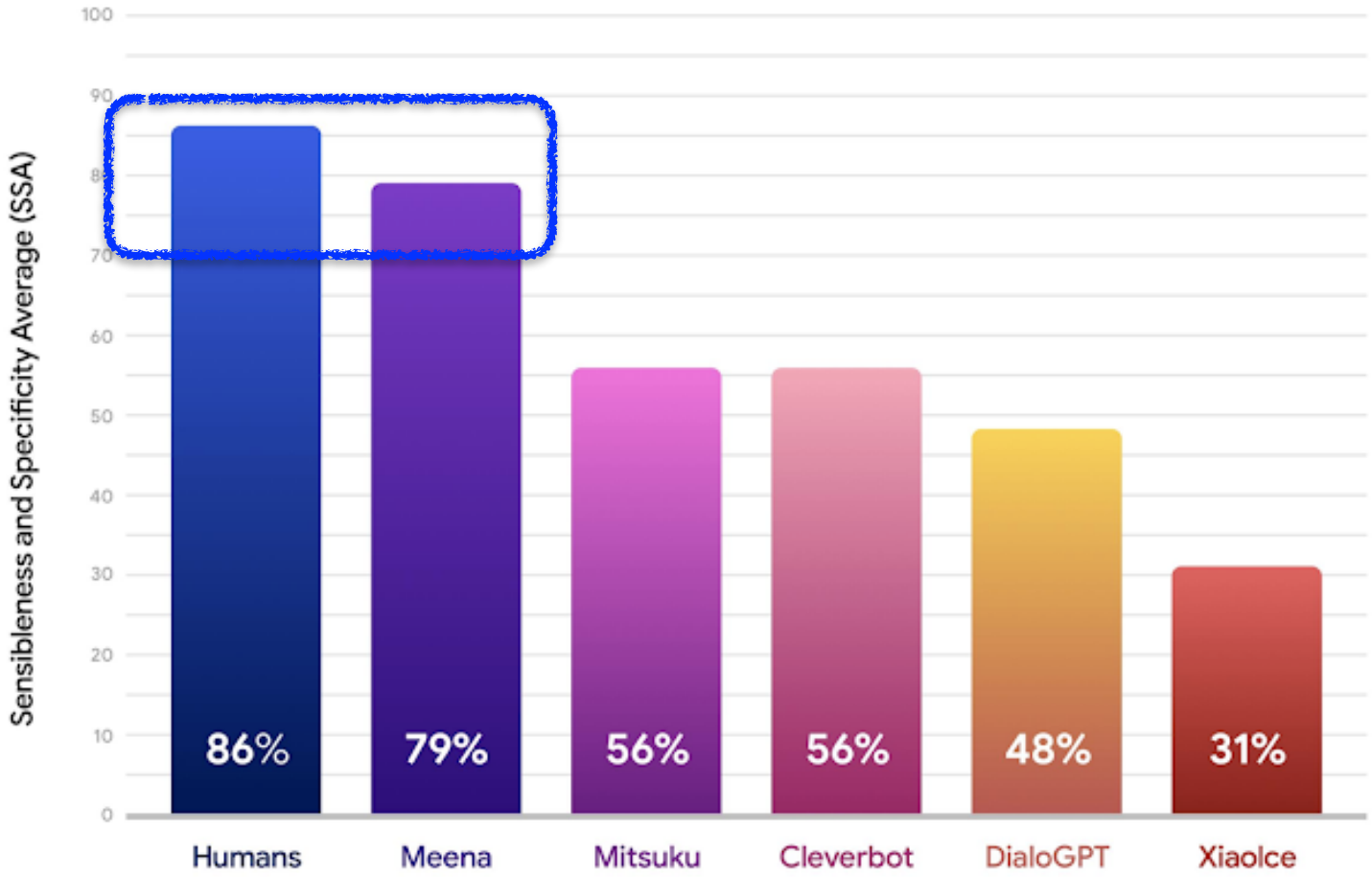
wikipedia.com

How Good Are Recent Conversational Agents?



https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

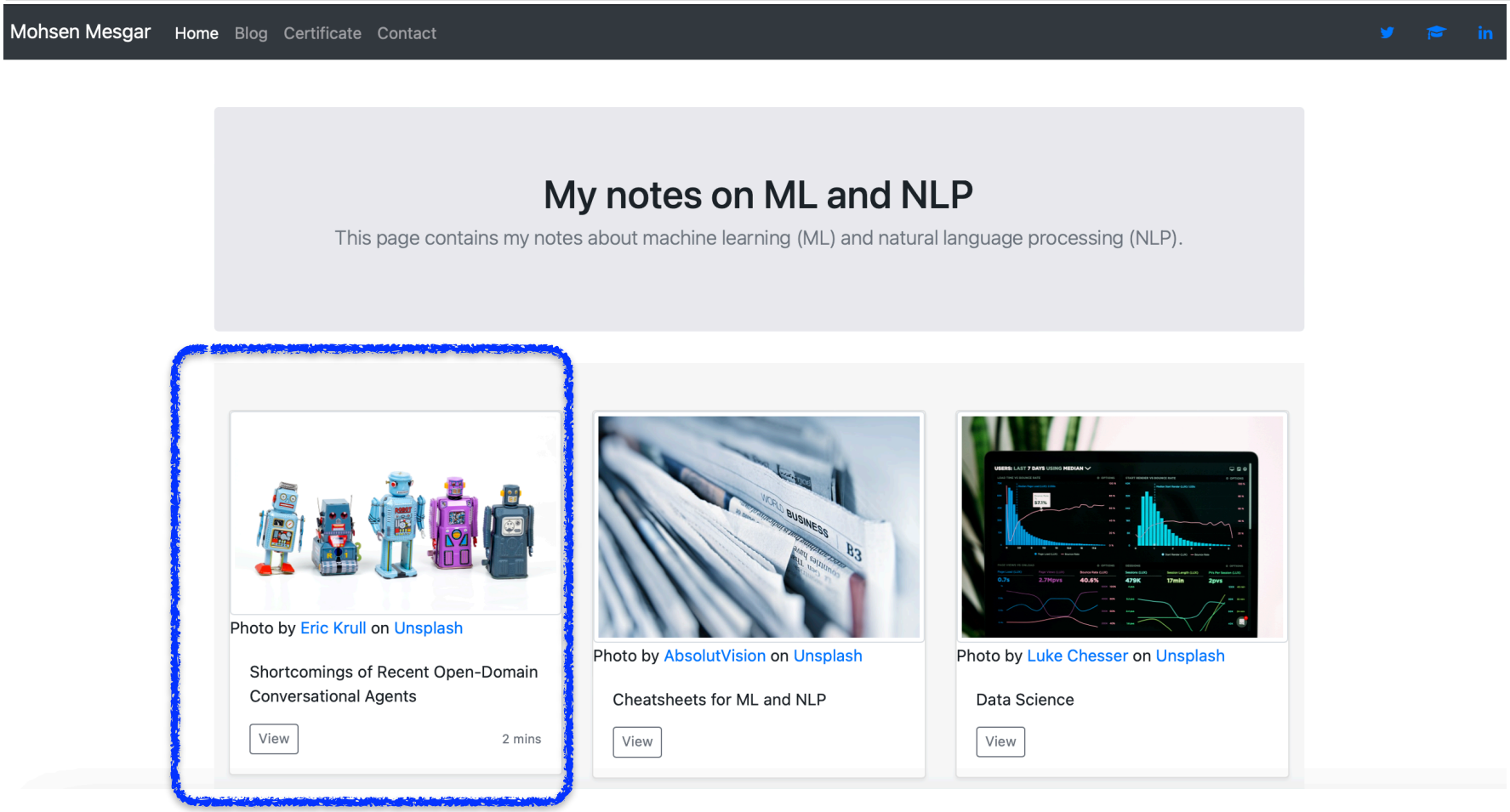
How Good Are Recent Conversational Agents?



https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

What Are Shortcomings?

TECHNISCHE UNIVERSITÄT DARMSTADT



What Are Shortcomings?

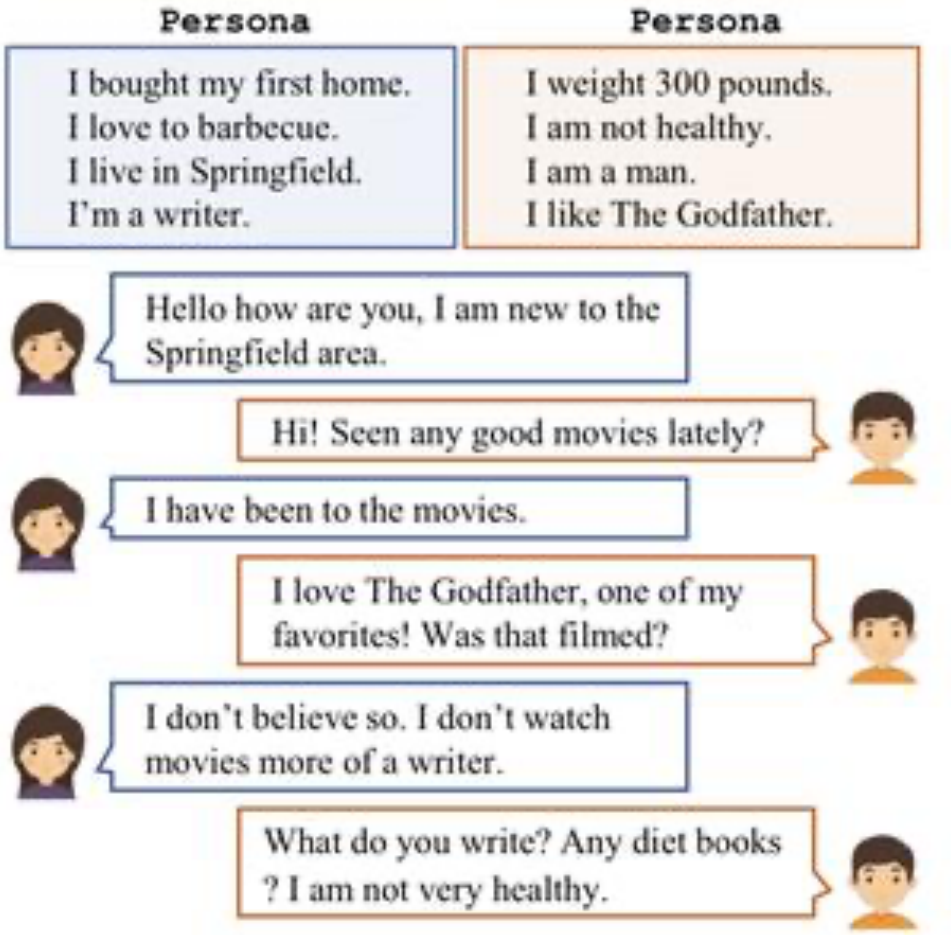
their intelligence. I categorize them into three groups.

1. Reliability

This group of shortcomings introduces how reliable current dialogue systems are in interactions with human users. This group includes: (1.1) the lack of factual consistency through conversations, (1.2) the lack of in-depth knowledge and inference in responses, (1.3) the lack of specificity and engaging use of knowledge, (1.4) the lacks of controllability and explainability, and (1.5) the lack of persona consistency to acquire users' trust and gain their long-term confidence.

2. Efficiency

Persona Consistency



Zhang et al., (ACL'18)

Persona Consistency

TECHNISCHE UNIVERSITÄT DARMSTADT

- Responses should ideally be
  - **semantically plausible**
    - **topically coherent**
    - **linguistically fluent**
  - **factually consistent with persona facts**

An Example Persona

TECHNISCHE UNIVERSITÄT DARMSTADT

Persona

Fact 1: i hate my job

Fact 2: i am 40 years old

Fact 3: i work as a car salesman

Fact 4: i am planning on getting a divorce

Fact 5: my wife spends all my money

U2

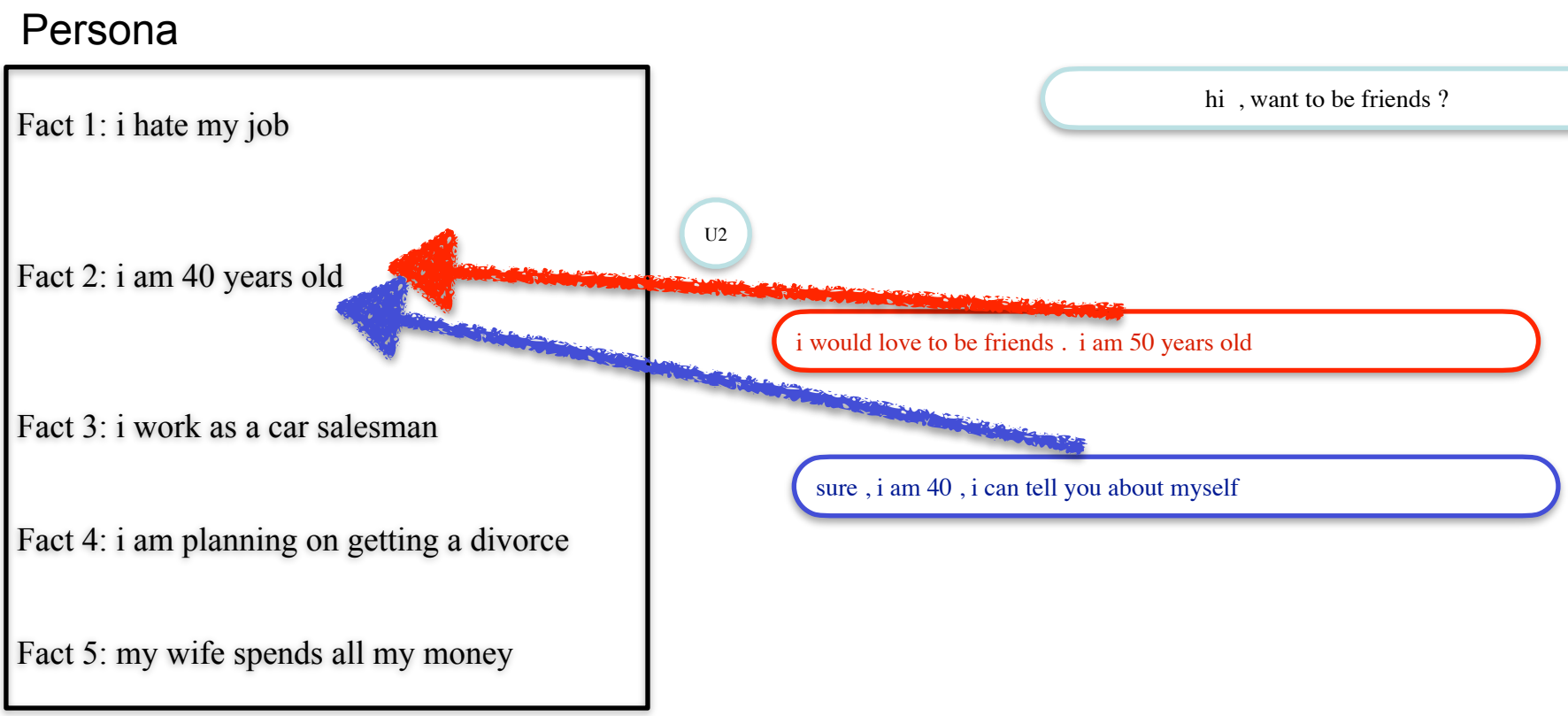
i would love to be friends . i am 50 years old

hi , want to be friends ?

U1

UKP

An Example Persona

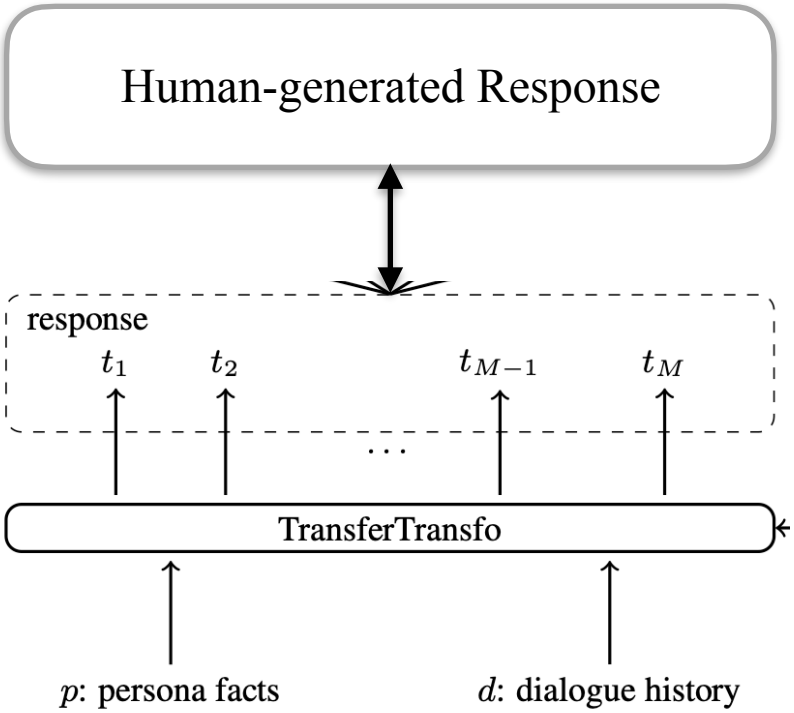


Our Goal

TECHNISCHE UNIVERSITÄT DARMSTADT

- We aim to **improve the response quality** in terms of
  - **factual consistency with facts about the given speaker's persona**
  - while **retaining its semantic plausibility**

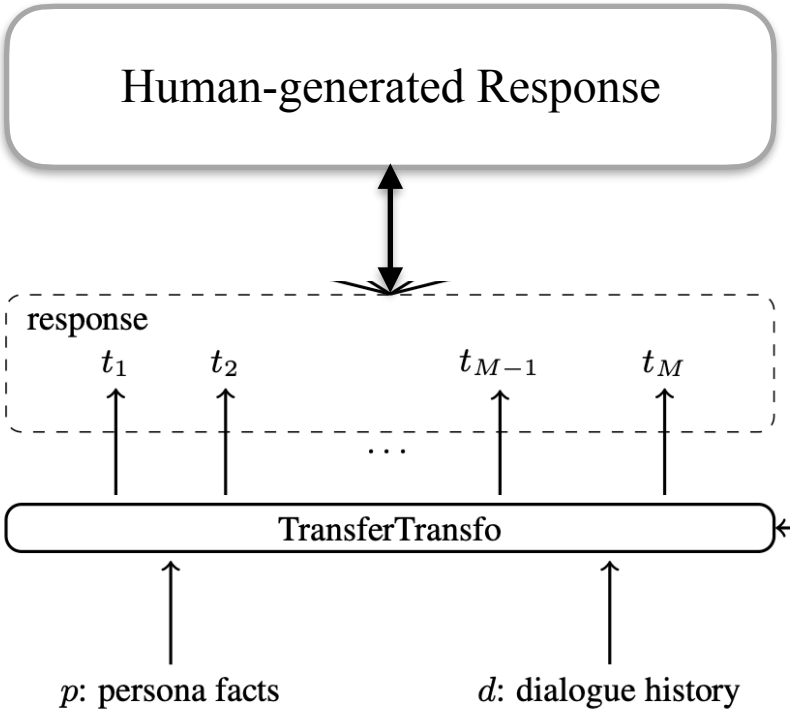
Related Work: TransferTransfo



Wolf et al, (ArXiv'19)

Related Work: TransferTransfo-SL

- **SL training objective** maximize the likelihood of human-written responses



Wolf et al, (ArXiv'19)

Related Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- While **SL has improved performanc**e, there **is still a misalignment between this training objective** – maximizing the likelihood of human-written responses – and **what we care about** –
  - **generating semantically plausible**
  - **factually consistent responses**

Related Work

TECHNISCHE UNIVERSITÄT DARMSTADT

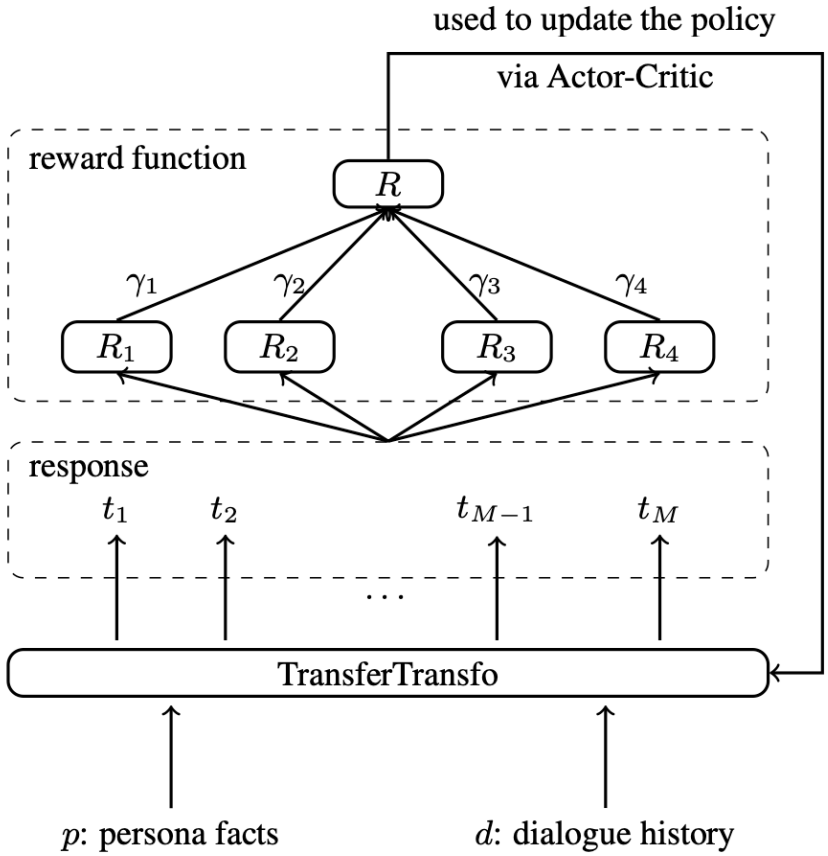
- This **misalignment has several reasons**:
  - The **maximum likelihood objective** considers **no distinction between primary errors** (e.g. inconsistent responses) and **unimportant errors** (e.g. selecting the precise word from a set of synonyms);
  - models **are incentivized to place probability mass on all human-generated responses**, including those that are low-quality;

Idea

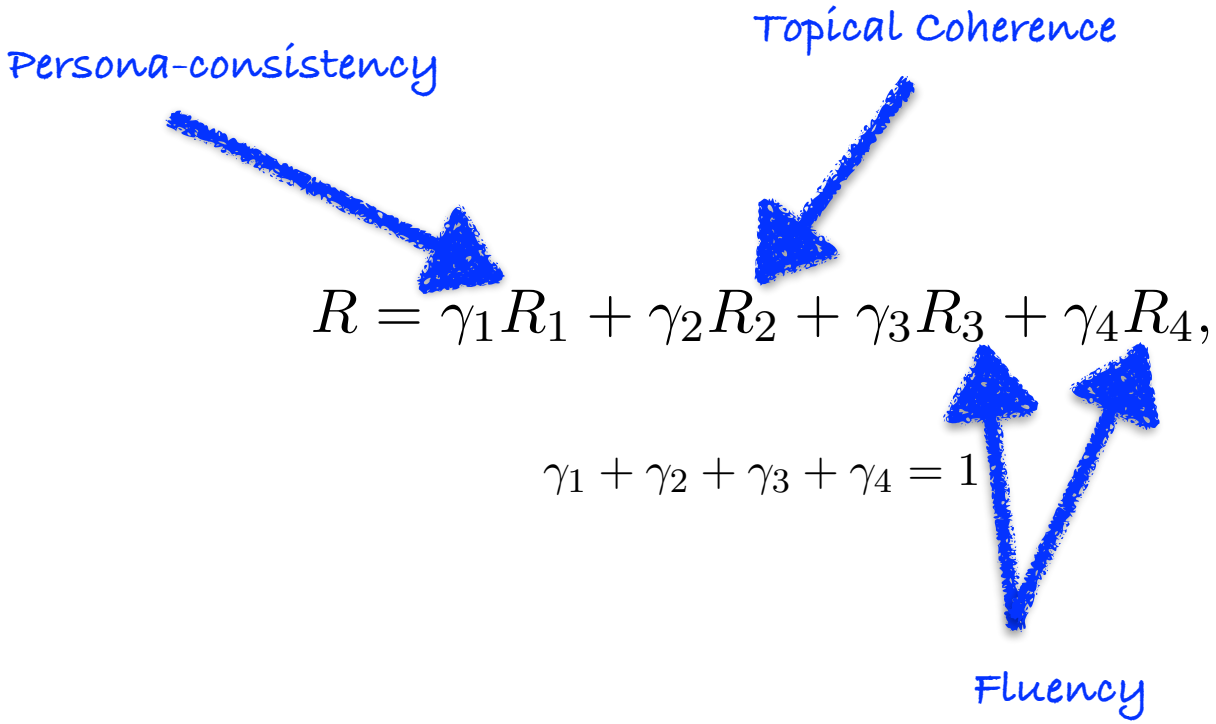
TECHNISCHE UNIVERSITÄT DARMSTADT

- Using **Reinforcement Learning (RL) with a multi-objective reward function** to take care of
  - **factual consistency with persona facts,**
  - **topical coherence with dialogue history,** and
  - **language fluency.**

Method: TransferTransfo-RL

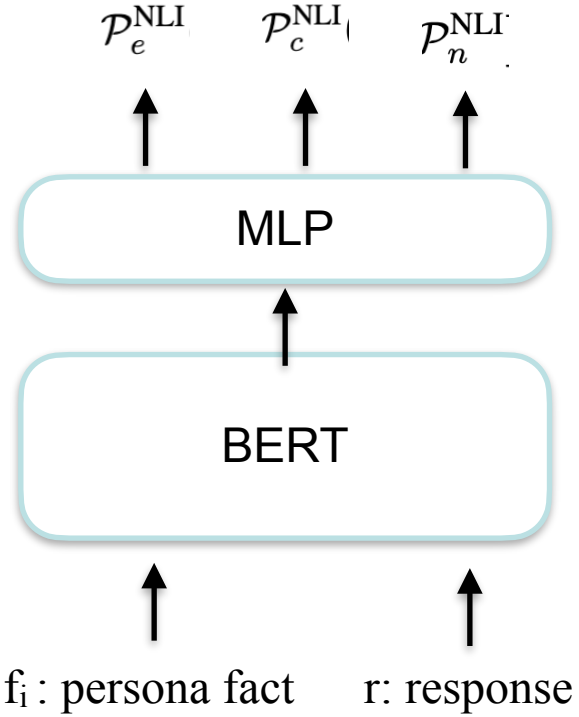


Reward


Persona-consistency
Topical Coherence
$R = \gamma_1 R_1 + \gamma_2 R_2 + \gamma_3 R_3 + \gamma_4 R_4,$
$\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1$
Fluency

Persona-consistency Sub-reward (R1)

Consistency with factual information, such as persona facts, can be characterized as a natural language inference (NLI) problem (Welleck et al., ACL'19; Dziri et al., NAACL'19)



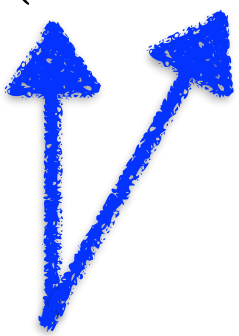
$$R_1 = \frac{1}{|p|}\sum_{f_i \in p} \mathcal{P}_e^{\text{NLI}}(f_i, r) - \frac{\beta}{|p|}\sum_{f_i \in p} \mathcal{P}_c^{\text{NLI}}(f_i, r),$$

Bert-based NLI model

Topical Coherence Sub-reward (R2)

TECHNISCHE UNIVERSITÄT DARMSTADT

$$R_2 = \cos(\vec{r}, \vec{u}_{T-1}).$$



Bert-based encoder

Fluency Sub-rewards (R3 and R4)

TECHNISCHE UNIVERSITÄT DARMSTADT

The content should be expressed fluently

$$R_3 = \frac{\alpha - \mathrm{NLL}(r)}{\alpha},$$

Fine-tuned Language Model

Repeated words should be avoided (See et al, ACL'19)

$$R_4 = 1 - \frac{\#\text{repeated-tokens-in-response}}{\#\text{tokens-in-response}}.$$

PersonaChat Corpus

TECHNISCHE UNIVERSITÄT DARMSTADT

- Consists of dialogues, in **English**, with 6 to 8 turns between randomly paired human crowd-workers.
- The workers were assigned **short text facts representing personas** and instructed **to talk to their dialogue partner *naturally to discover each other's persona*.**
- We chose this corpus because of **its focus on promoting natural conversations** while **grounding conversations in the persona facts**.
- Each persona consists of 4 or 5 facts, and on average is assigned to 8.3 unique dialogues.

	Train	Validation
Num. of dialogues	17,878	1,000
Num. of utterances	262,876	15,602
Num. of personas	955	200

Results

TECHNISCHE UNIVERSITÄT DARMSTADT

Method	PPL	F1	BLEU	PC
TransferTransfo-SL	21.31	17.06	0.065	09.32
TransferTransfo-RL	22.64	**17.78**	**0.067**	**13.06**

Number of entailments

Number of contradictions

$$\text{PC} = 100\frac{N_e - N_c}{N},$$

Number of response-fact pairs in the evaluation set

Results - Looking at PC in Detail

TECHNISCHE UNIVERSITÄT DARMSTADT

	Consistent	Contradicting	Neutral
Automatic Evaluation			
TransferTransfo-SL	11.14	01.82	87.04
TransferTransfo-RL	**14.81**	**01.75**	**83.43**
Δ	3.41 $\uparrow$	0.07 $\downarrow$	3.61 $\downarrow$
Human Evaluation			
TransferTransfo-SL	43.71	17.71	38.58
TransferTransfo-RL	**52.71**	**14.00**	**33.29**
Δ	9.00 $\uparrow$	3.71 $\downarrow$	5.29 $\downarrow$

Semantic Plausibility: Human Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

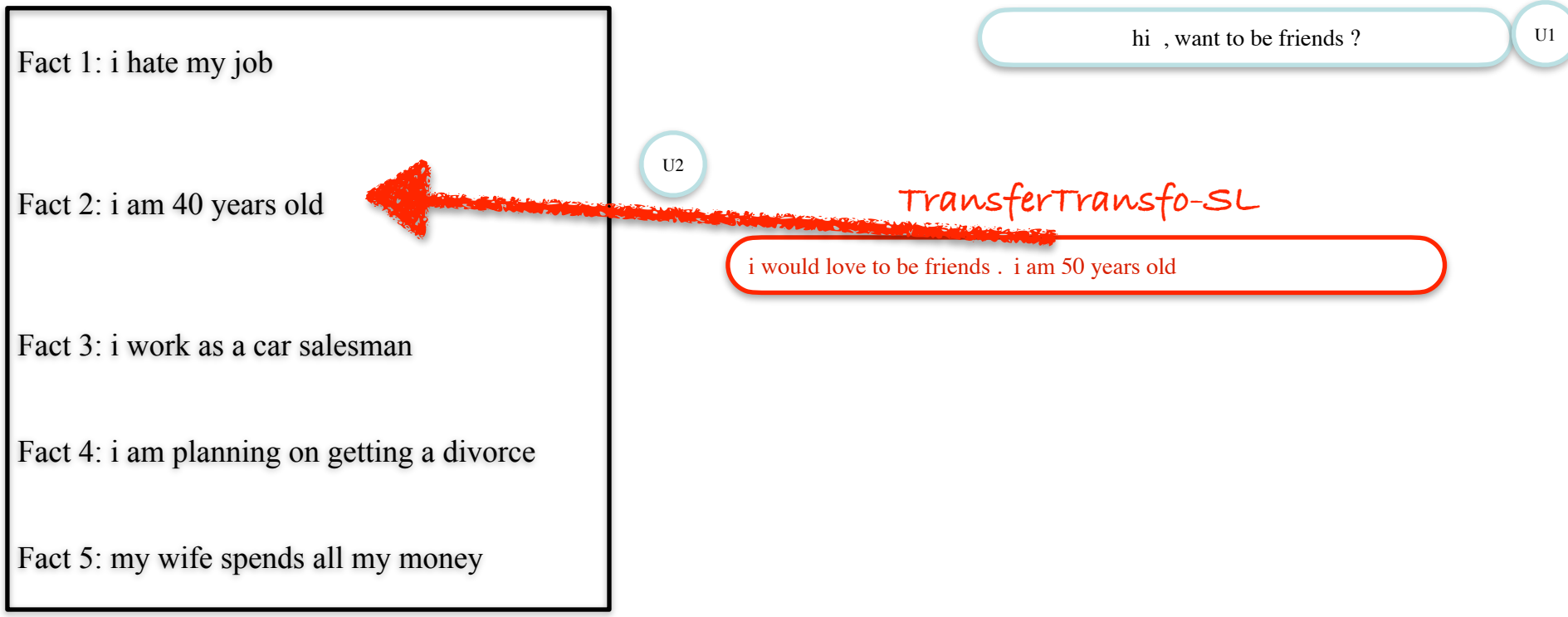
Method	Average Semantic Plausibility
TransferTransfo-SL	3.33
TransferTransfo-RL	**3.50**

Table 5: Human evaluation: semantic plausibility.

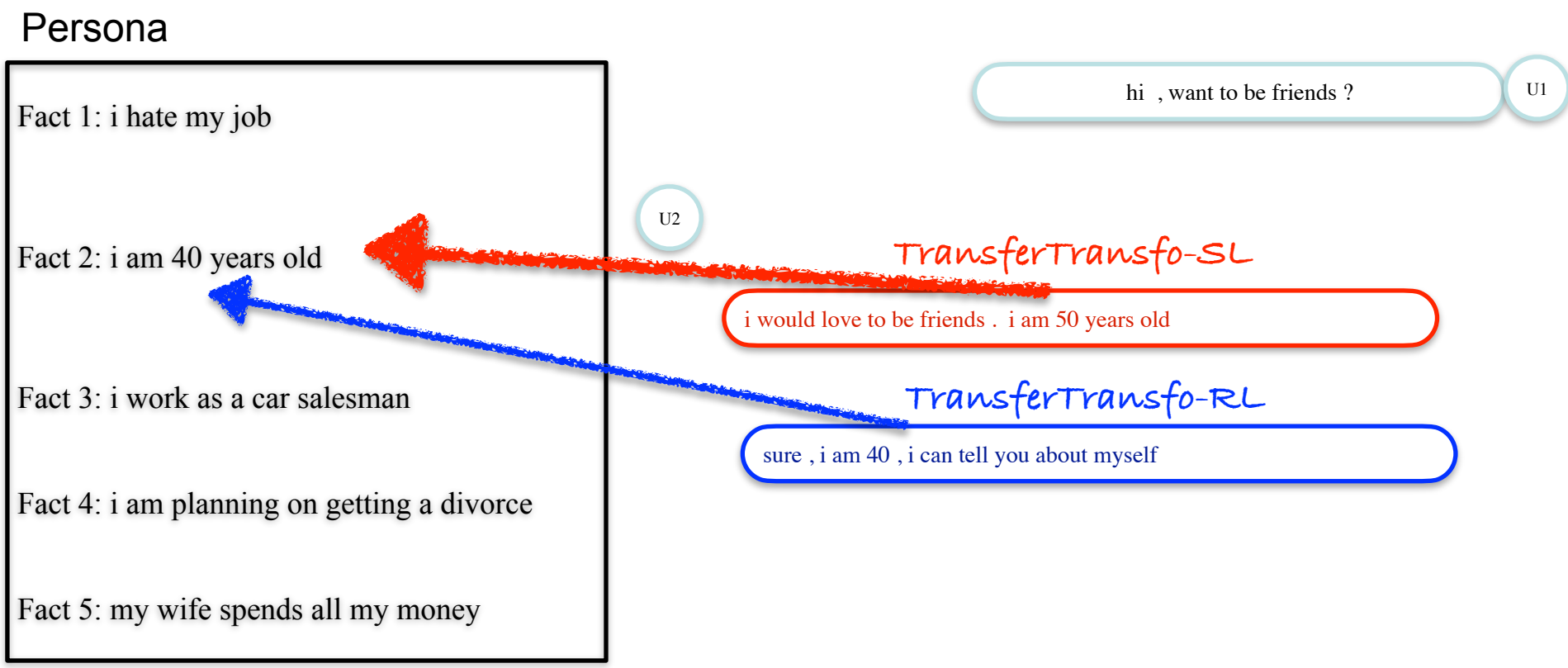
We also ask the human judges to rate the semantic plausibility of each response with an ordinal score ranging from 1 (worst) to 5 (best), encompassing

- *coherence*,
- *grammatical correctness*, and
- *low repetitiveness*.

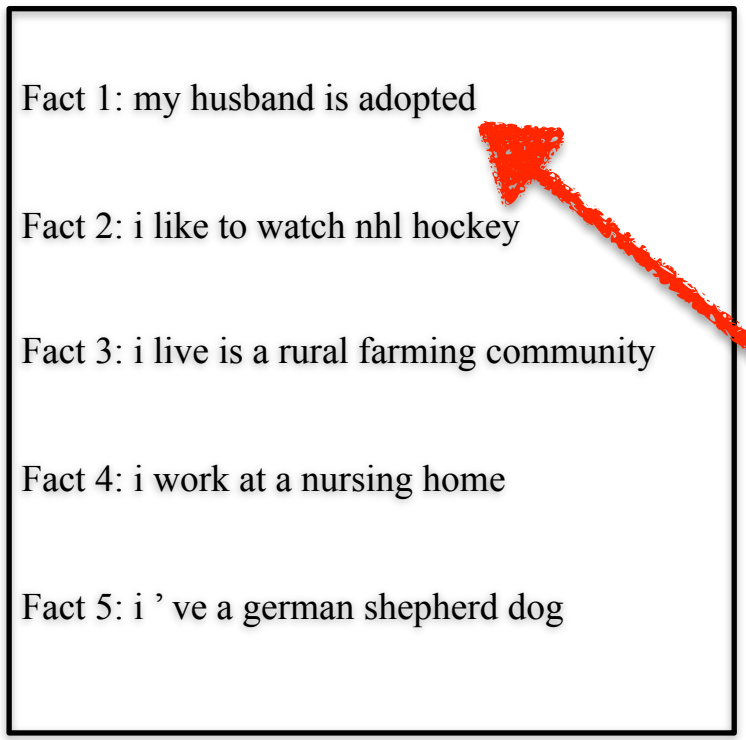
Example


Persona
Fact 1: i hate my job
Fact 2: i am 40 years old
Fact 3: i work as a car salesman
Fact 4: i am planning on getting a divorce
Fact 5: my wife spends all my money
U2
hi , want to be friends ?
U1
TransferTransfo-SL
i would love to be friends . i am 50 years old

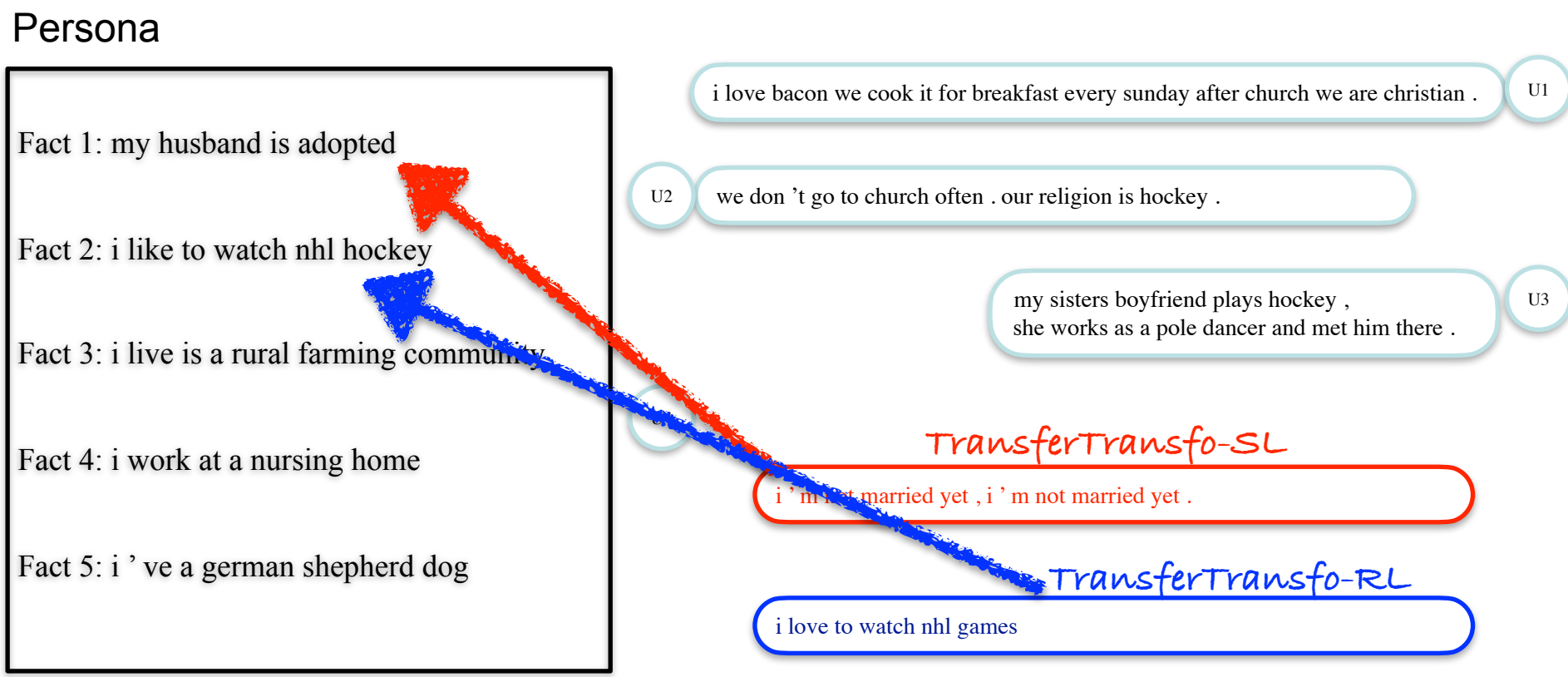
Example


Persona
Fact 1: i hate my job
Fact 2: i am 40 years old
Fact 3: i work as a car salesman
Fact 4: i am planning on getting a divorce
Fact 5: my wife spends all my money
hi , want to be friends ?
U1
U2
TransferTransfo-SL
i would love to be friends . i am 50 years old
TransferTransfo-RL
sure , i am 40 , i can tell you about myself

Example


Persona
Fact 1: my husband is adopted
Fact 2: i like to watch nhl hockey
Fact 3: i live is a rural farming community
Fact 4: i work at a nursing home
Fact 5: i ' ve a german shepherd dog
i love bacon we cook it for breakfast every sunday after church we are christian .
U1
U2
we don 't go to church often . our religion is hockey .
my sisters boyfriend plays hockey ,
she works as a pole dancer and met him there .
U3
U4
TransferTransfo-SL
i ' m not married yet , i ' m not married yet .

Example


Persona
Fact 1: my husband is adopted
Fact 2: i like to watch nhl hockey
Fact 3: i live is a rural farming community
Fact 4: i work at a nursing home
Fact 5: i ' ve a german shepherd dog
i love bacon we cook it for breakfast every sunday after church we are christian .
U1
U2
we don 't go to church often . our religion is hockey .
my sisters boyfriend plays hockey ,
she works as a pole dancer and met him there .
U3
TransferTransfo-SL
i ' m not married yet , i ' m not married yet .
TransferTransfo-RL
i love to watch nhl games

Privacy and Fairness

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Personalized AI agents memorize a lot of personalized data. Agents should preserve the privacy of such data and** do not reveal sensitive information in their responses **to unauthorized questions and adversarial attacks**
- AI agents learn from **a lot of training data**. Such data might represent **social biases** in favor and/or against different **genders**, **religions**, **languages**, **under-represented** people, etc. However, having access to user's persona information shouldn't make **AI agents to generate outputs reflecting social biases**

New and interesting problems are ahead …

Privacy and Fairness

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Privacy and Fairness

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Privacy and Fairness

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

Thank You