Alexa, can you help me?

I don't know what to do.

Dialog Systems

João Sedoc

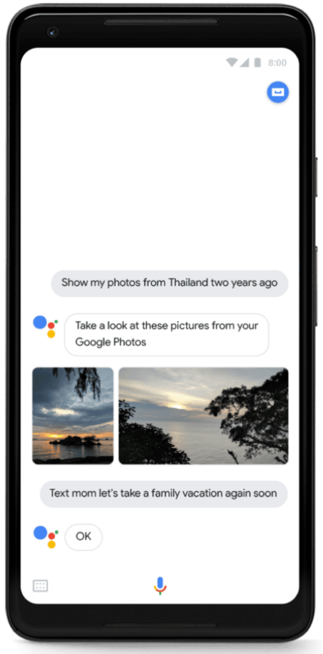
jsedoc@nyu.edu

New York University

Department of Technology, Operations, and Statistics

Chatbots are Ubiquitous: Personal Agents, Games, Education, Business & Medicine

Woebot



Lots of Tools

amazon LEX

chatfuel

Dialogflow

flow.ai

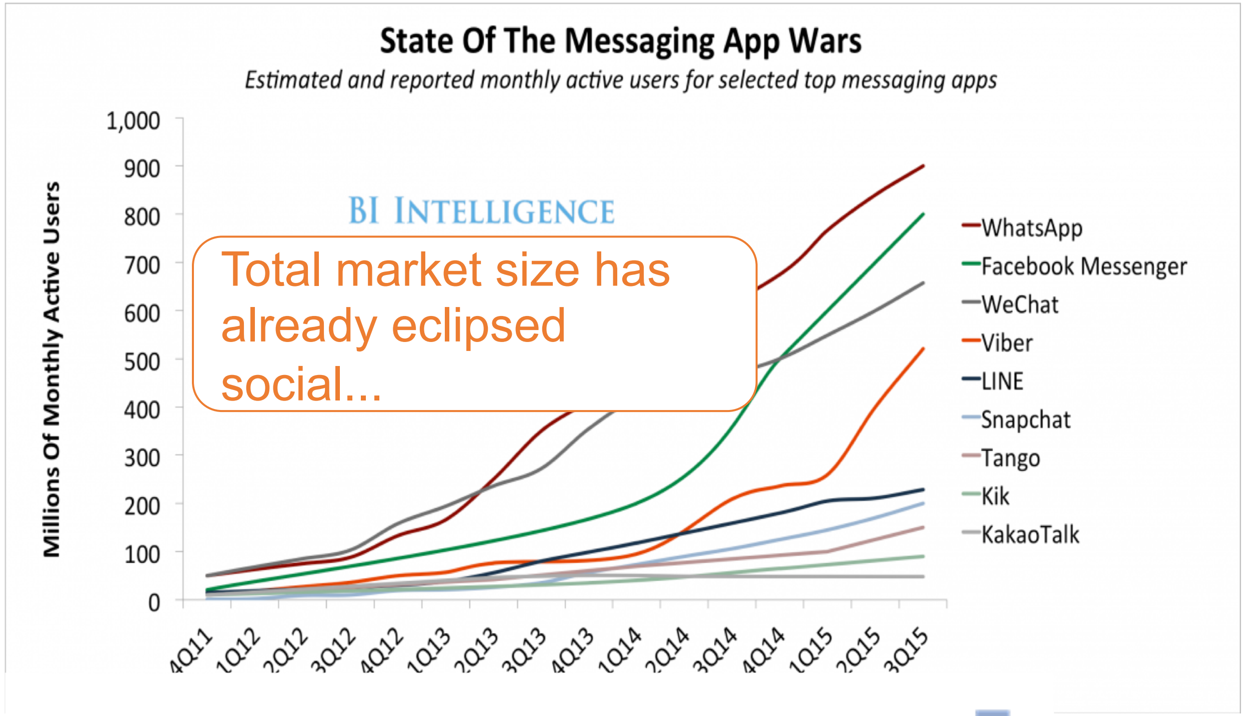
RASA

Watson Assistant

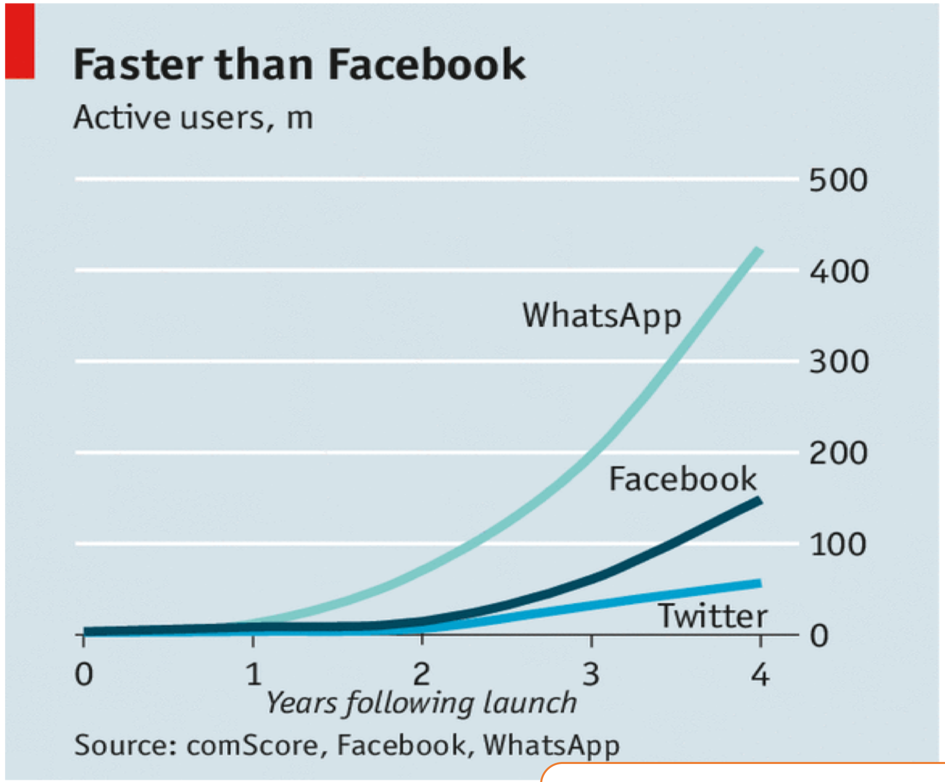
KITT·AI

https://docs.google.com/spreadsheets/d/1RgG-dRS42EHlG7QdJOTg2ZO587KutTTPeUfyxVKoIn8/edit#gid=0

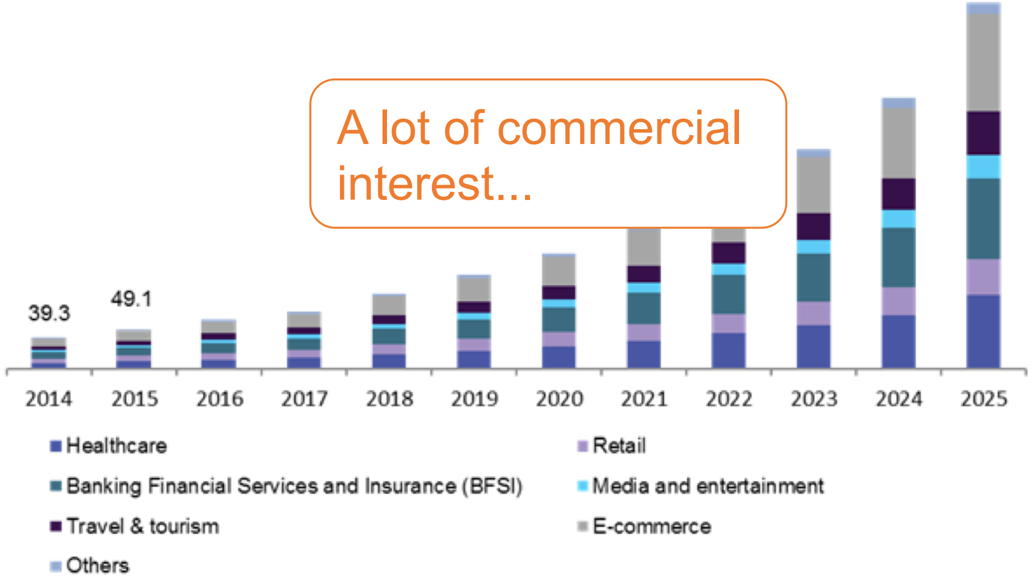
Motivation - Lots of Chat



Total market size has already eclipsed social...



Adoption has been faster than social...



A lot of commercial interest...

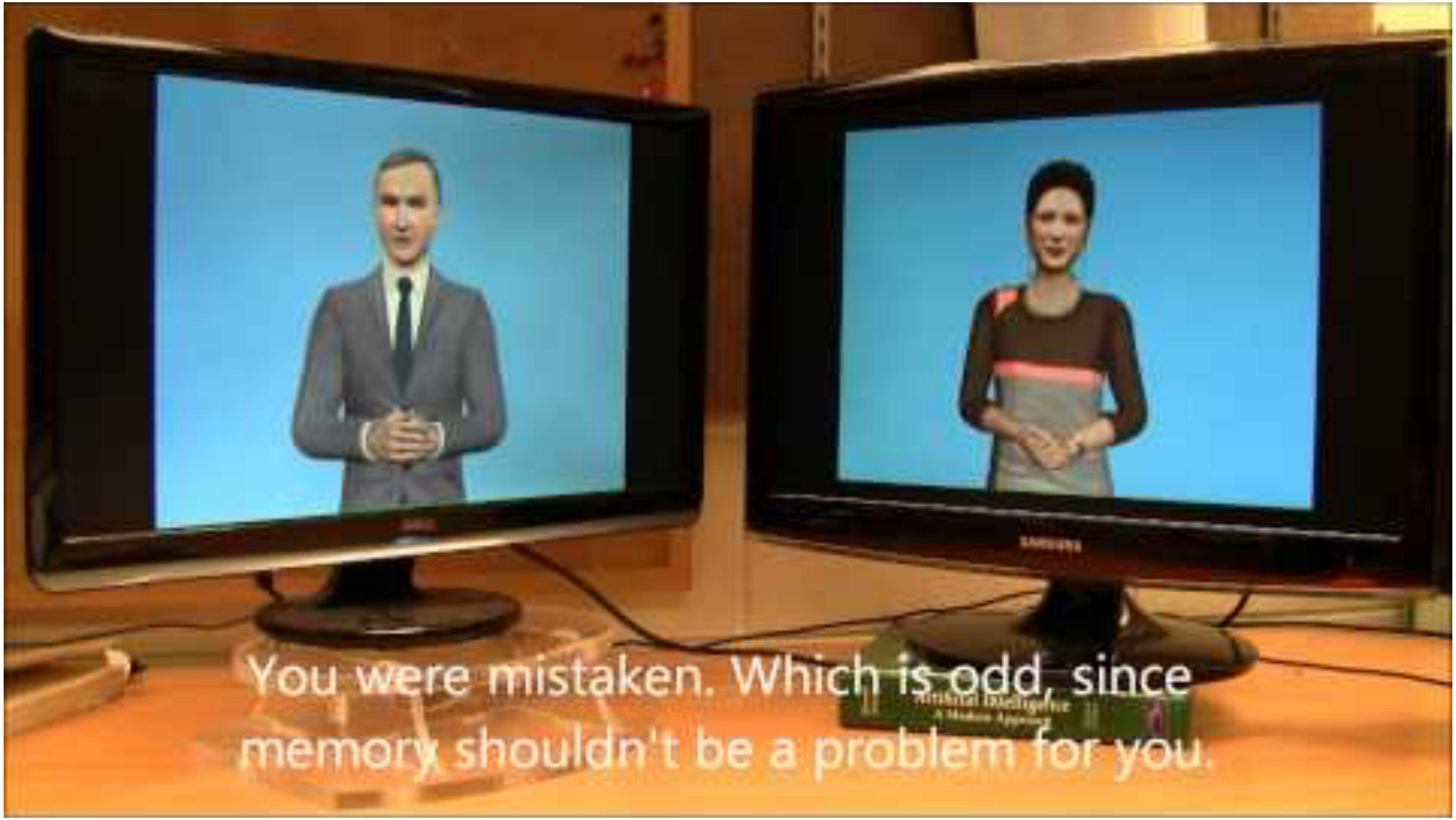
Artificial Intelligence

- Can robots understand language?
- Can robots actually think?
- Not clear definition of intelligence or how to measure it!

- The Turing Test (1950)
- Indirect assessment of intelligent behaviour

(Image adapted from: http://www.clubic.com/mag/culture/actualite-751397-imitation-game-alan-turing-pere-informatique.html)

AI with AI conversations: Cleverbot (Carpenter, 2011)



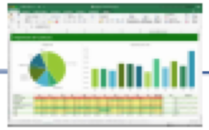
Challenges for Artificial Intelligence

- **Knowledge Representation**
  - ❑ about learning, storing and retrieving relevant infor-mation about the world and one's previous experiences
- **Commonsense reasoning***
  - ❑ about using world knowledge for interpreting, explaining and predicting daily life events and outcomes

Aspirational Goal: Enterprise Assistant

Where are sales lagging behind our forecast?

Task Completion



The worst region is [country], where sales are XX% below projections

Do you know why?

QA (decision support)

The forecast for [product] growth was overly optimistic

How can we turn this around?

Here are the 10 customers in [country] with the most growth potential, per our CRM model

Can you set up a meeting with the CTO of [company]?

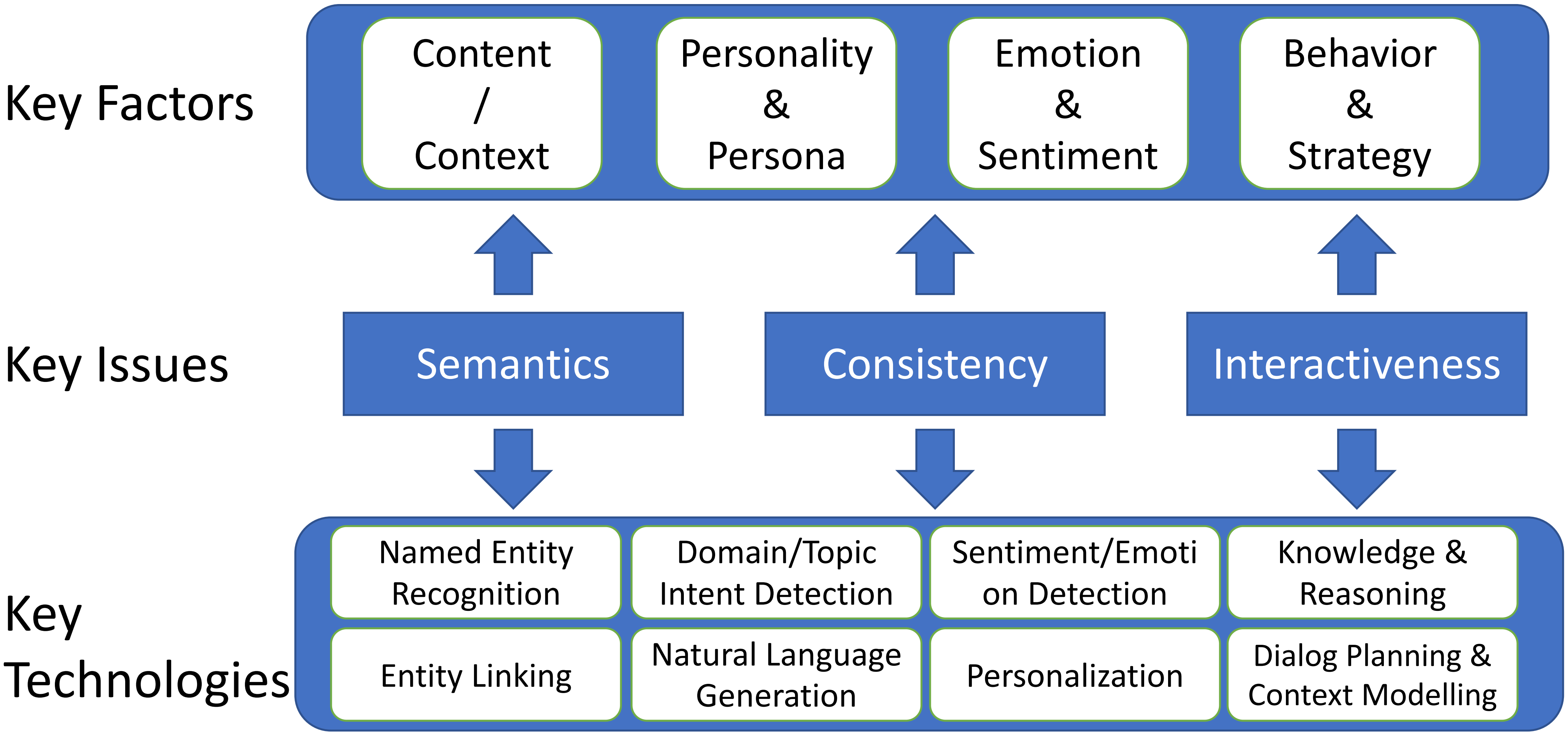
Info Consumption

Task Completion

Yes, I've set up a meeting with [person name] for next month when you're in [location]

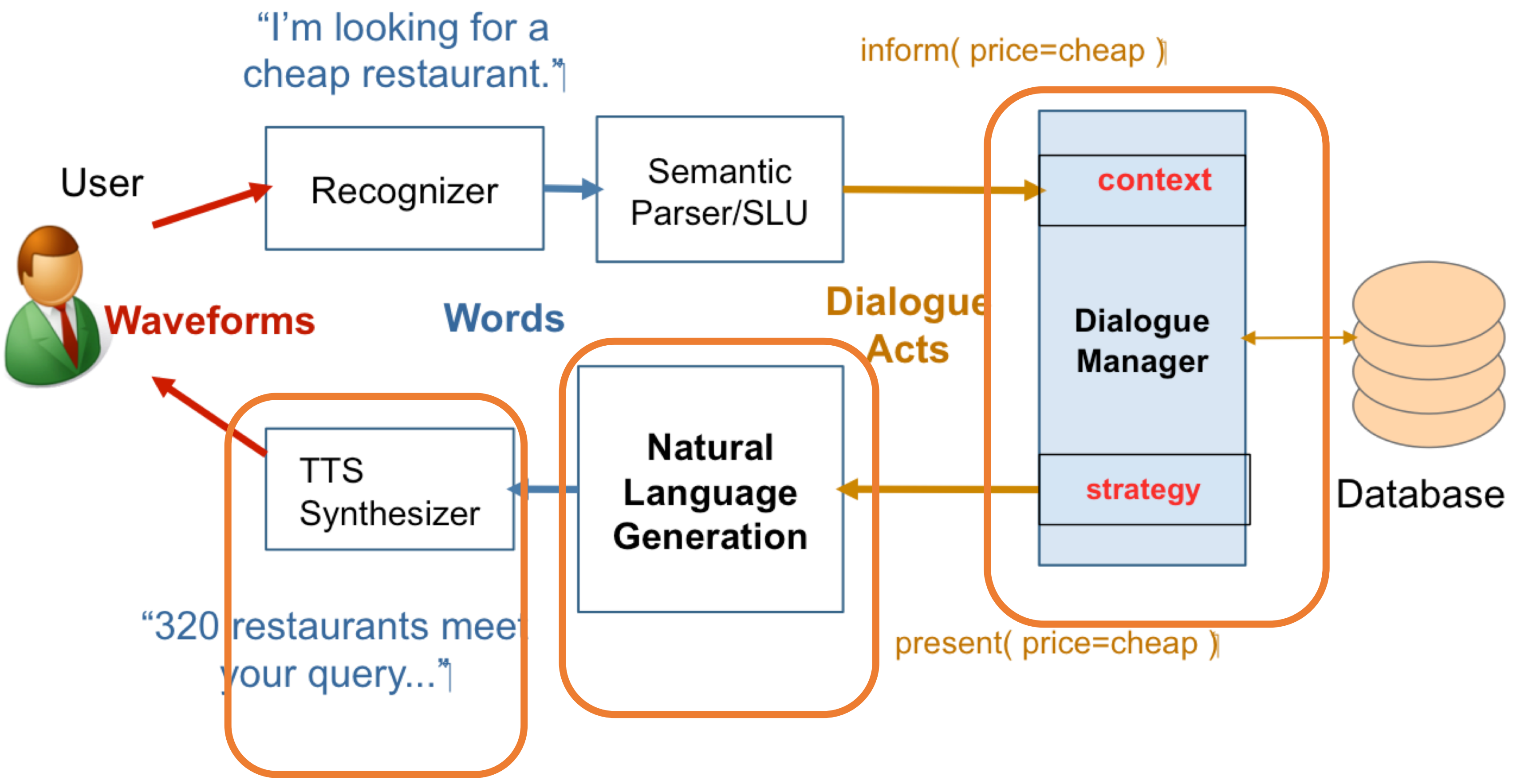
Thanks

Challenges for Conversational Agents



From Huang et al., 2019, “Challenges in Building Intelligent Open-Domain Systems”

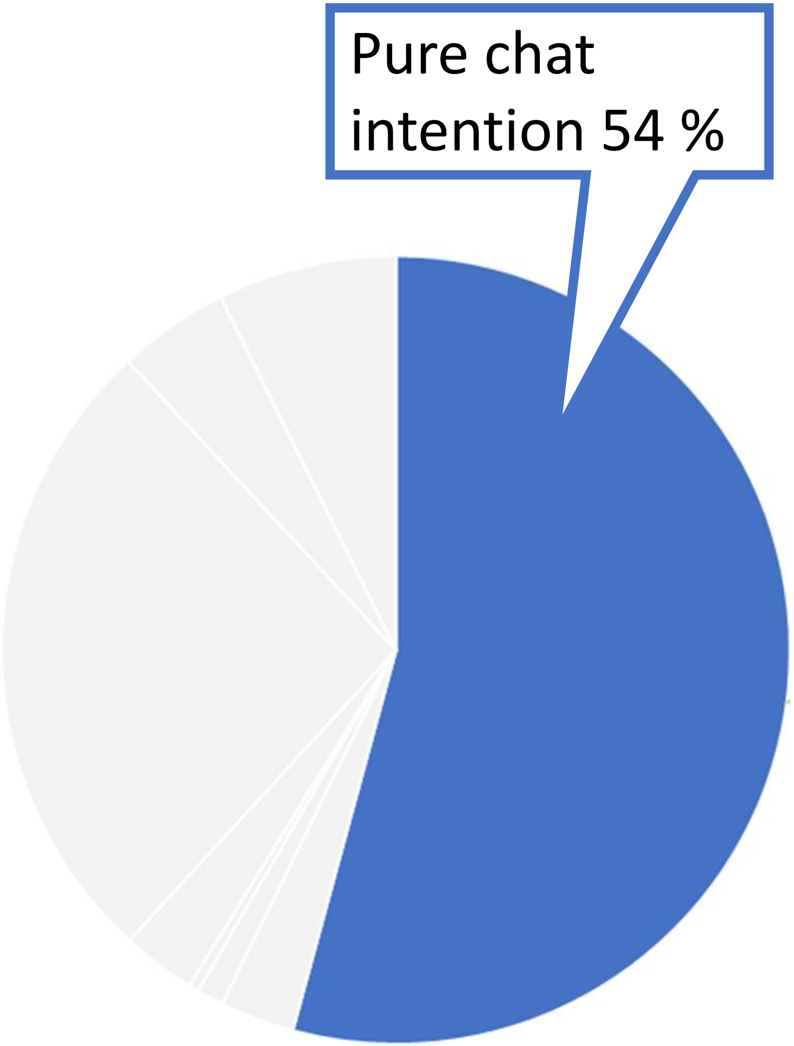
Spoke Dialog System Architecture



Two Types of Systems

1. Chatbots
2. Goal-based (Dialog agents)
   - SIRI, interfaces to cars, robots, …
   - Booking flights, restaurants, or question answering

Social Chat is Natural in Dialog Systems



Analysis is done from 100 sessions randomly sampled from the log of Microsoft Rinna, a commercial chatbot with 6M+ users in Japan.

Source: Wu & Yan, Deep Chit-Chat: Deep Learning for ChatBots Tutorial EMNLP 2018

Chatbot Architectures

Rule-based

1. Pattern-action rules (Eliza)
   \+ a mental model (Parry)

Corpus-based (from large chat corpus)

2. Information Retrieval
3. Neural network encoder-decoder

Eliza pattern/transform rules

`(0 YOU 0 ME)` [*pattern*]

→

`(WHAT MAKES YOU THINK I 3 YOU)`
[*transform*]

0 means Kleene *

The 3 is the constituent # in pattern

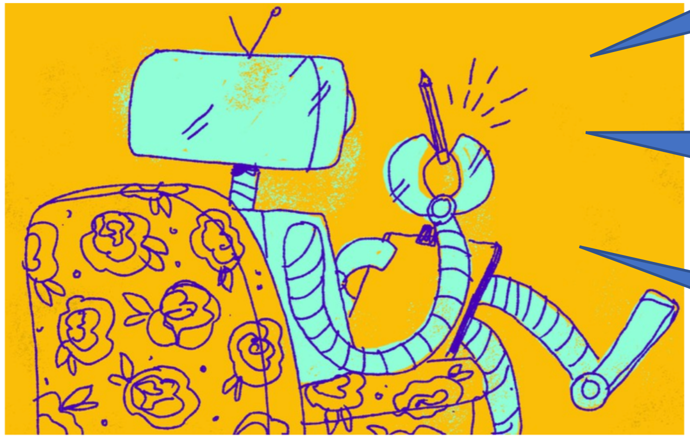
`You hate me`

`WHAT MAKES YOU THINK I HATE YOU`

http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm

Personality in chatbots: Eliza and Parry

Eliza



Parry

Good Evening. Tell me your problems.

People get on my nerves sometimes.

I am not sure I understand you fully.

You should pay more attention.

Suppose you should pay more attention.

You're entitled to your own opinion.

Parry's persona

- 28-year-old single man, post office clerk
- no siblings and lives alone
- sensitive about his physical appearance, his family, his religion, his education and the topic of sex.
- hobbies are movies and gambling on horseracing,
- recently attacked a bookie, claiming the bookie did not pay off in a bet.
- afterwards worried about possible underworld retaliation
- eager to tell his story to non-threating listeners.

Chatbot Architectures

Rule-based

1. Pattern-action rules (Eliza)

   + a mental model (Parry)

Corpus-based (from large chat corpus)

2. Information Retrieval
3. Neural network encoder-decoder

Information Retrieval based Chatbots

Idea: Mine conversations of human chats or human-machine chats

Microblogs: Twitter or Weibo (微博)

Movie dialogs

- Cleverbot (Carpenter 2017 http://www.cleverbot.com)
- Microsoft XiaoIce
- Microsoft Tay

Two IR-based Chatbot Architectures

1. Return the response to the most similar turn
   - Take user's turn (*q*) and find a (tf-idf) similar turn *t* in the corpus C

     q = "do you like Doctor Who"

     t' = "do you like Doctor Strangelove"
   - Grab whatever the response was to *t*.

$$r = response\left(\underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}\right)$$

Yes, so funny

2. Return the most similar turn

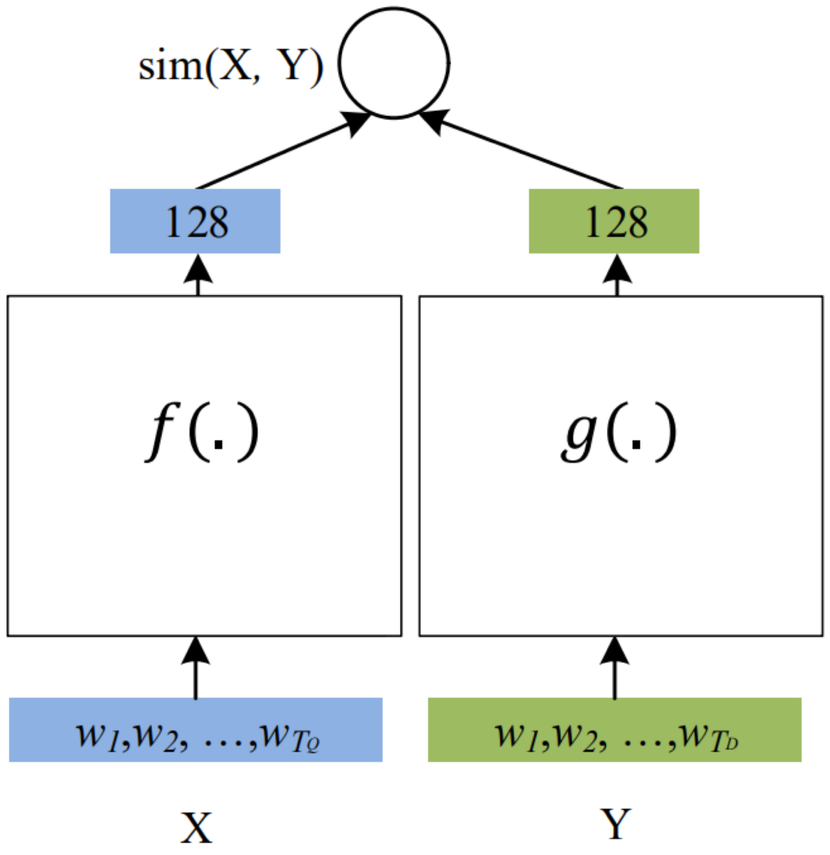
$$r = \underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}$$

Do you like Doctor Strangelove

Deep Semantic Similarity Model

Relevance measured by cosine similarity

Word sequence x_t



Learning: maximize the similarity between X (source) and Y (target)

Representation: use DNN to extract abstract semantic features, f or g is a

- Multi-Layer Perceptron (MLP) if text is a bag of words [Huang+ 13]
- **Convolutional Neural Network (CNN) if text is a bag of chunks** [Shen+ 14]
- Recurrent Neural Network (RNN) if text is a sequence of words [Palangi+ 16]

Chatbot Architectures

Rule-based

1. Pattern-action rules (Eliza)
   + a mental model (Parry)

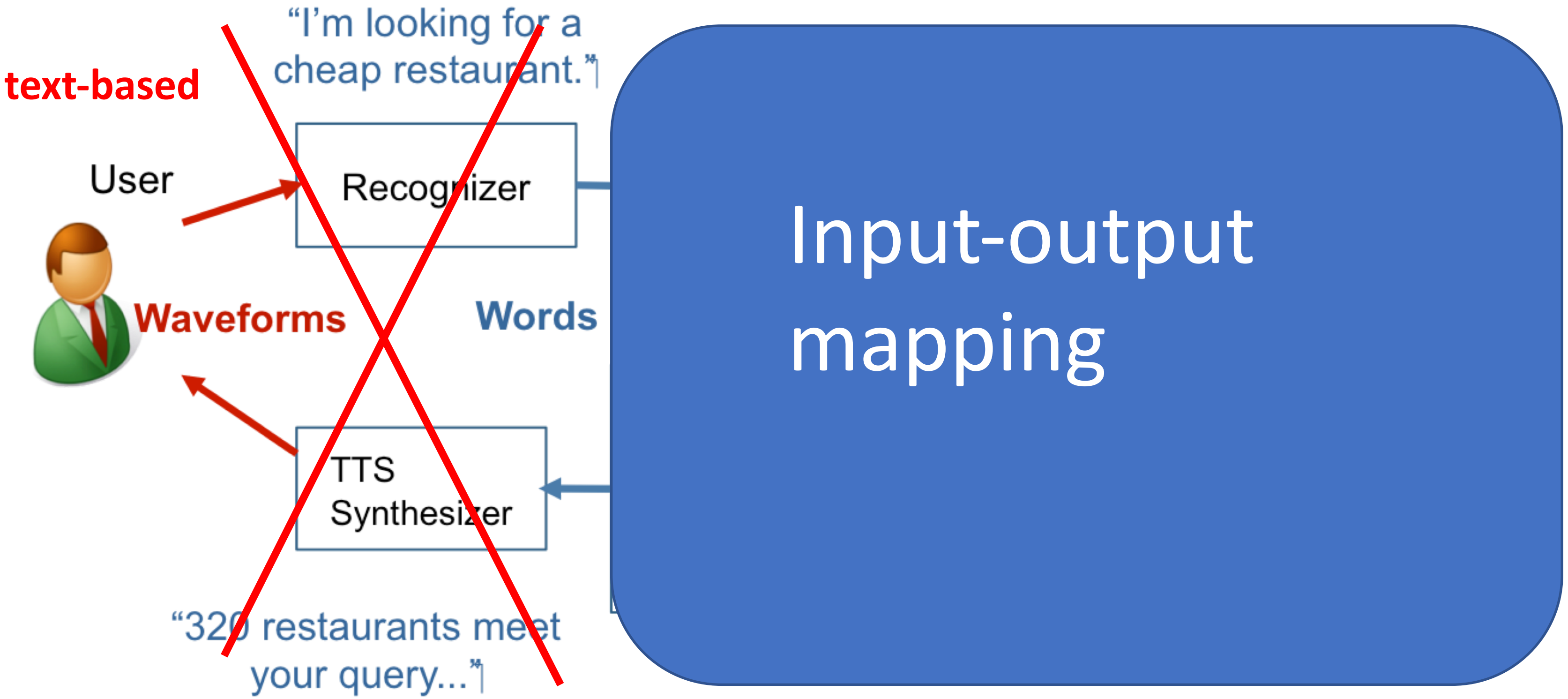
Corpus-based (from large chat corpus)

2. Information Retrieval
3. Neural network encoder-decoder

Neural Network Encoder-Decoder Generative Models

Response Generation Systems

- **End-to-end** systems.
- Learn from “raw” dialogue data (e.g. OpenSubtitles).
- No semantic or pragmatic annotation required.
- Mainly successful in open-domain, non-task oriented systems.



Neural Conversation Model (NCM)

vs

Rule-Based Model (Cleverbot)

User: are you a follower or a leader ?
CleverBot: no !!
NCM: i 'm a leader .

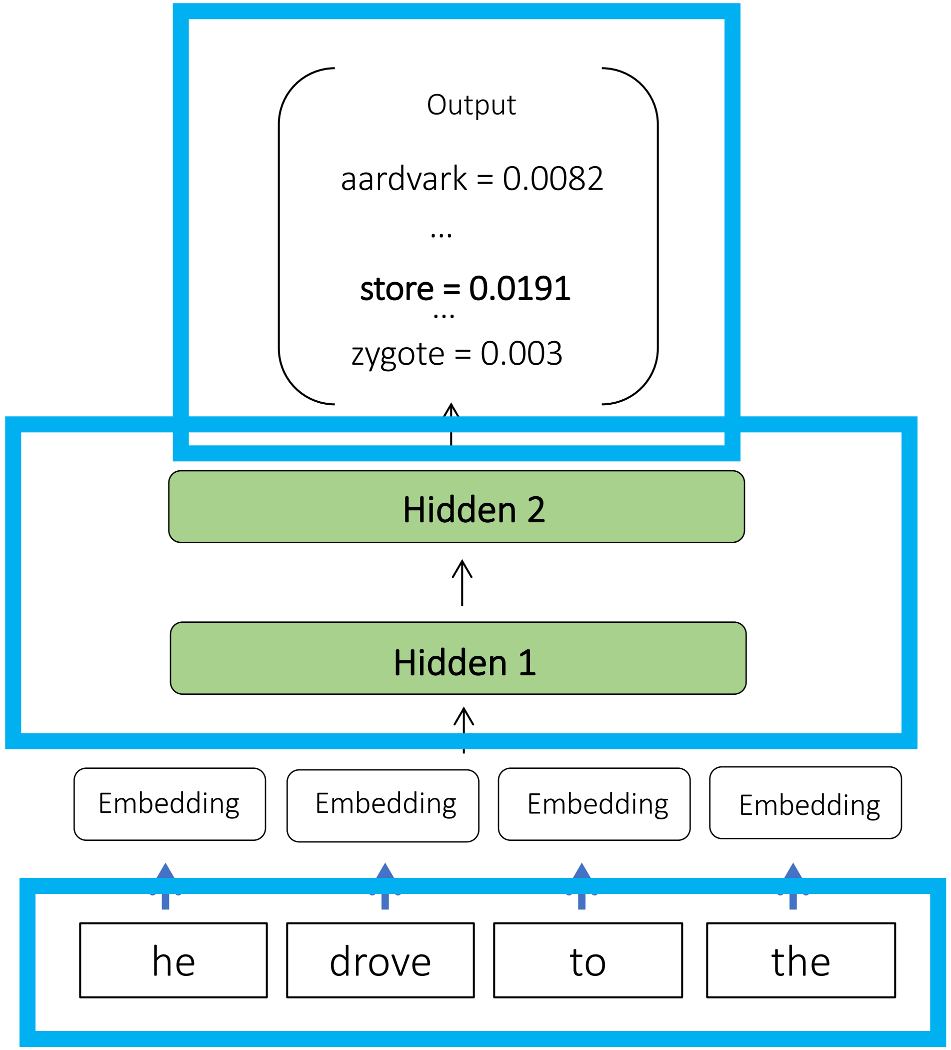
Vinyals and Le 2015

"A Neural Conversation Model"

Image borrowed from [farizrahman4u/seq2seq]()

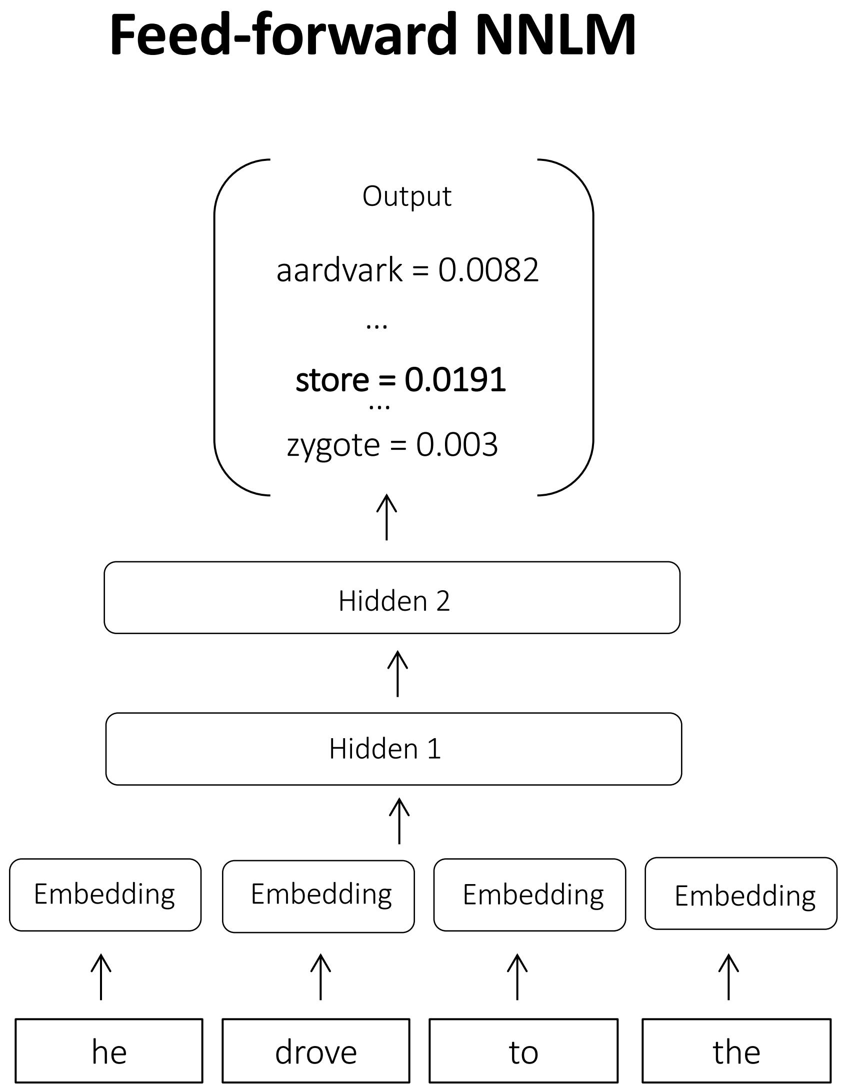
Neural Network Language Models (NNLMs)

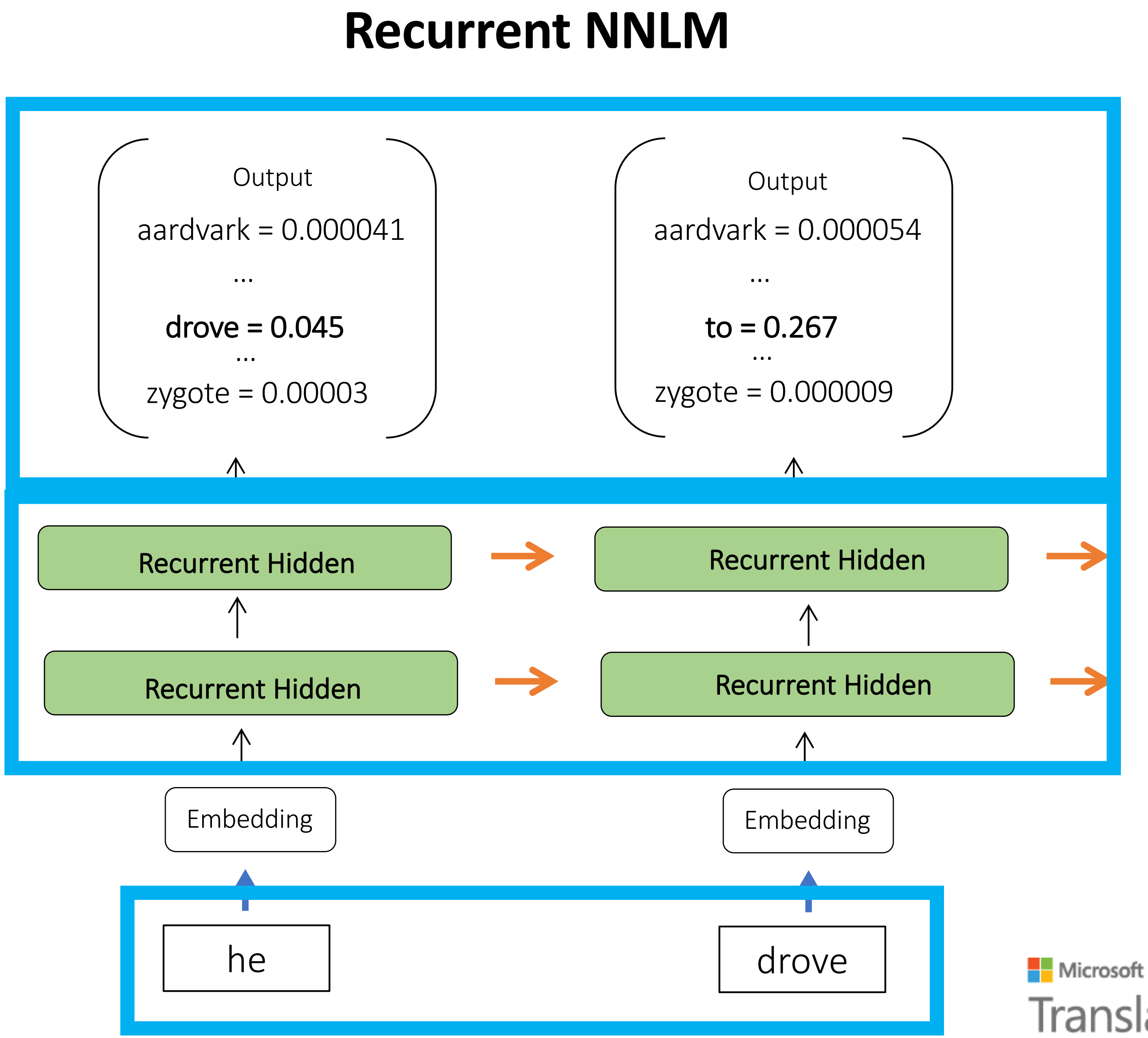
Feed-forward NNLM



Microsoft Translator

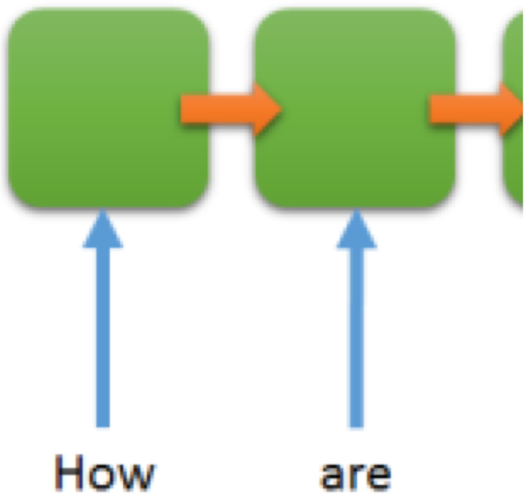
Neural Network Language Models (NNLMs)

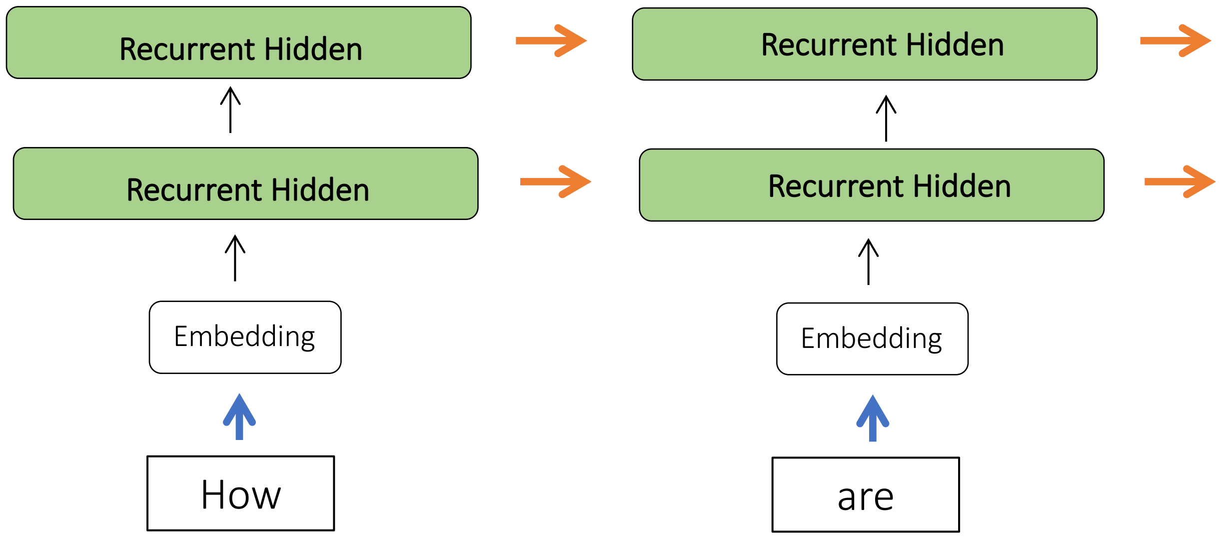




Microsoft Translator

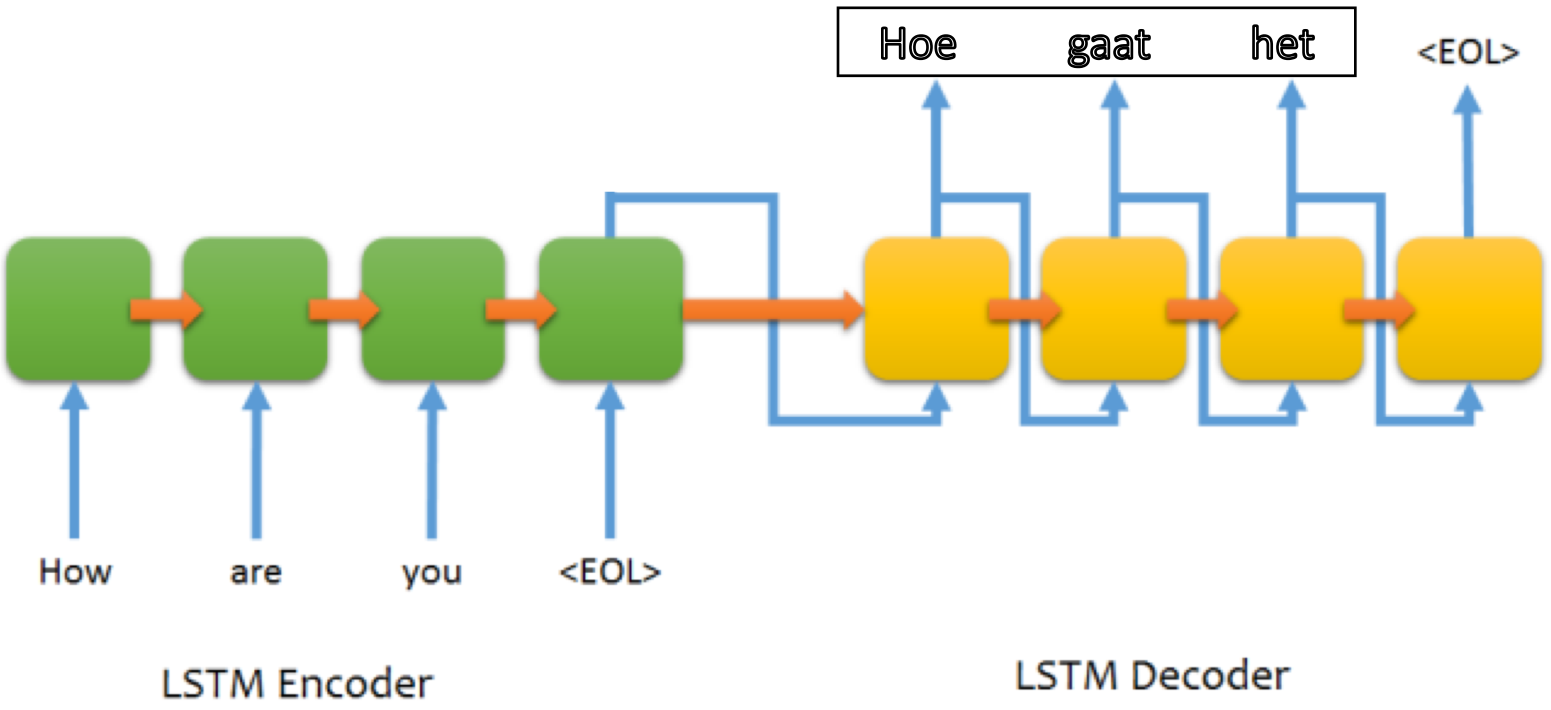
Sentence Encoder





Microsoft Translator

Sequence to Sequence Model

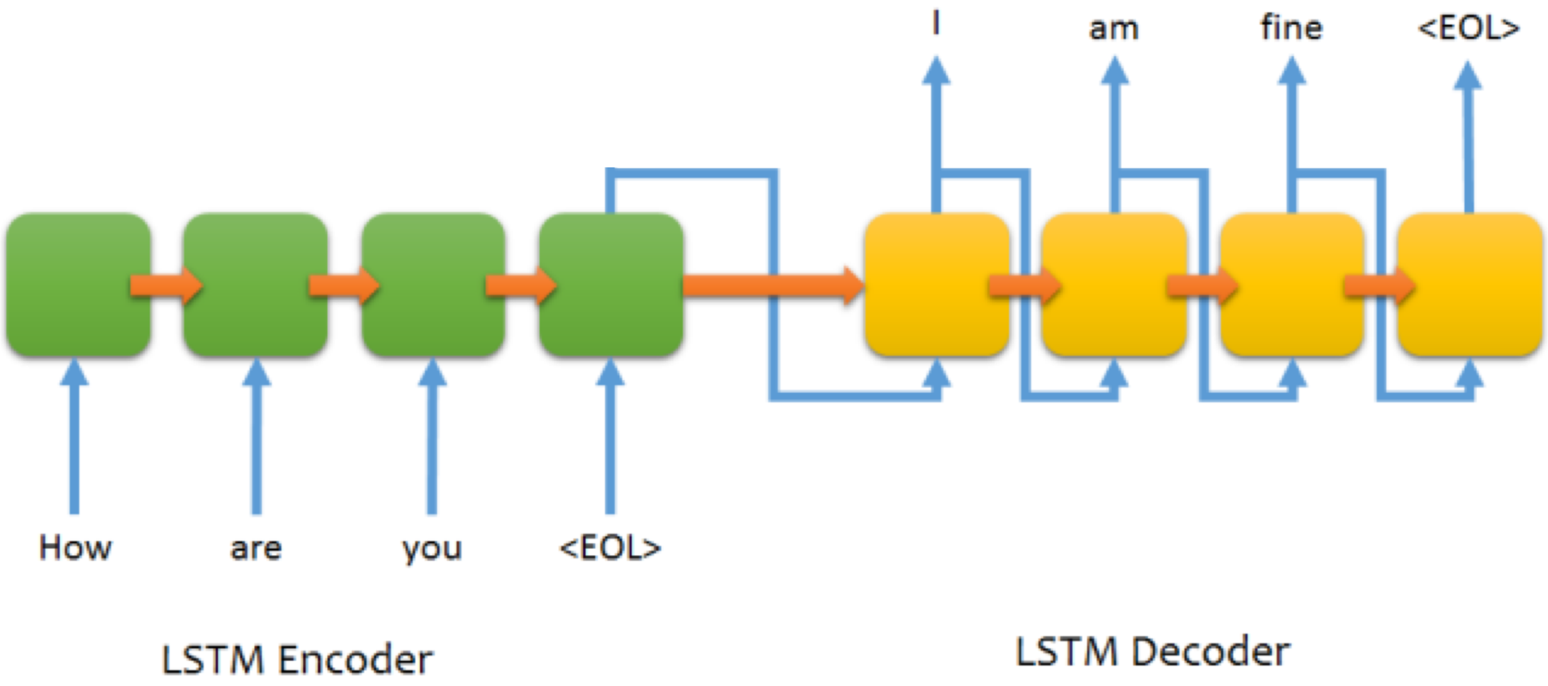


Sutskever et al. 2014

"Sequence to Sequence Learning with Neural Networks"

Image borrowed from [farizrahman4u/seq2seq]()

Sequence to Sequence Model

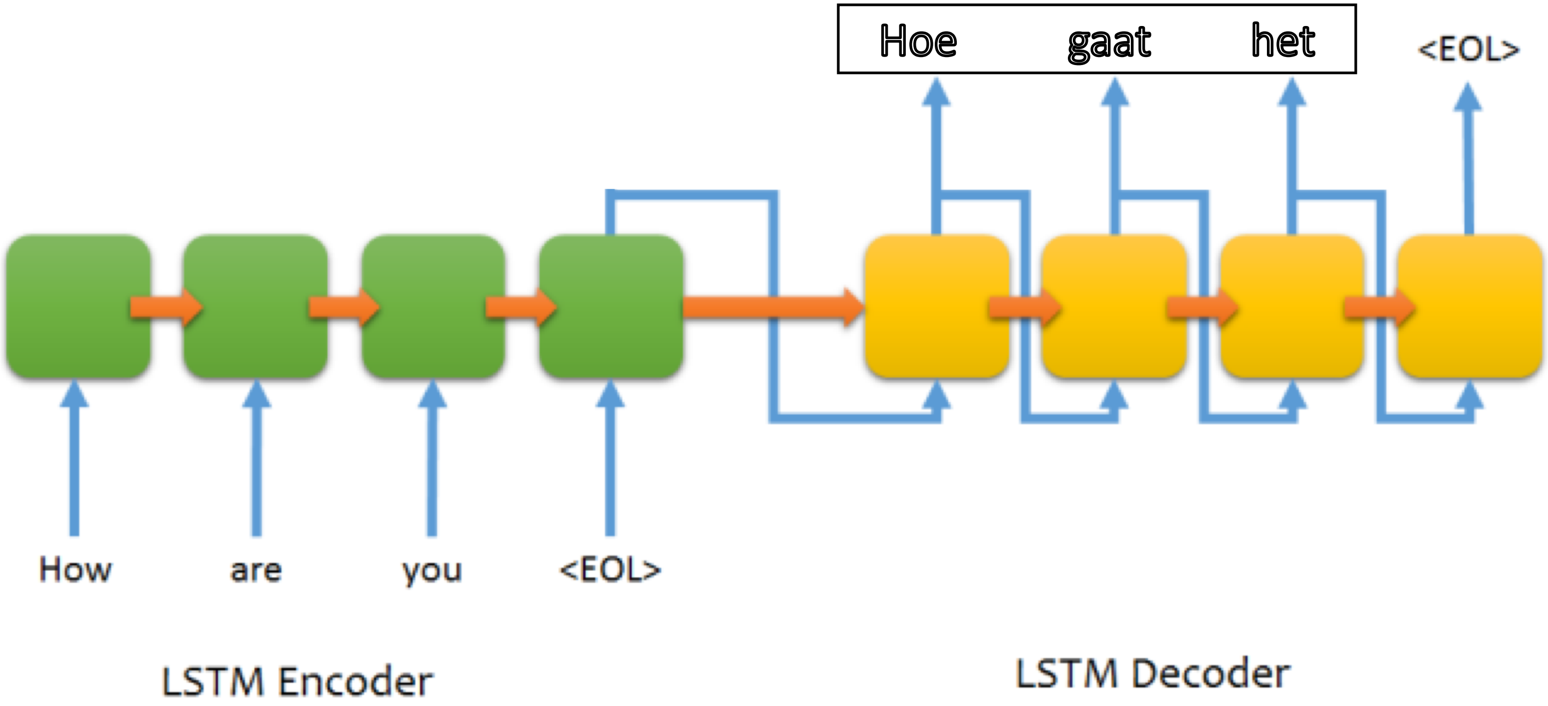


Vinyals and Le 2015

“A Neural Conversation Model”

Image borrowed from [farizrahman4u/seq2seq]

Sequence to Sequence Model

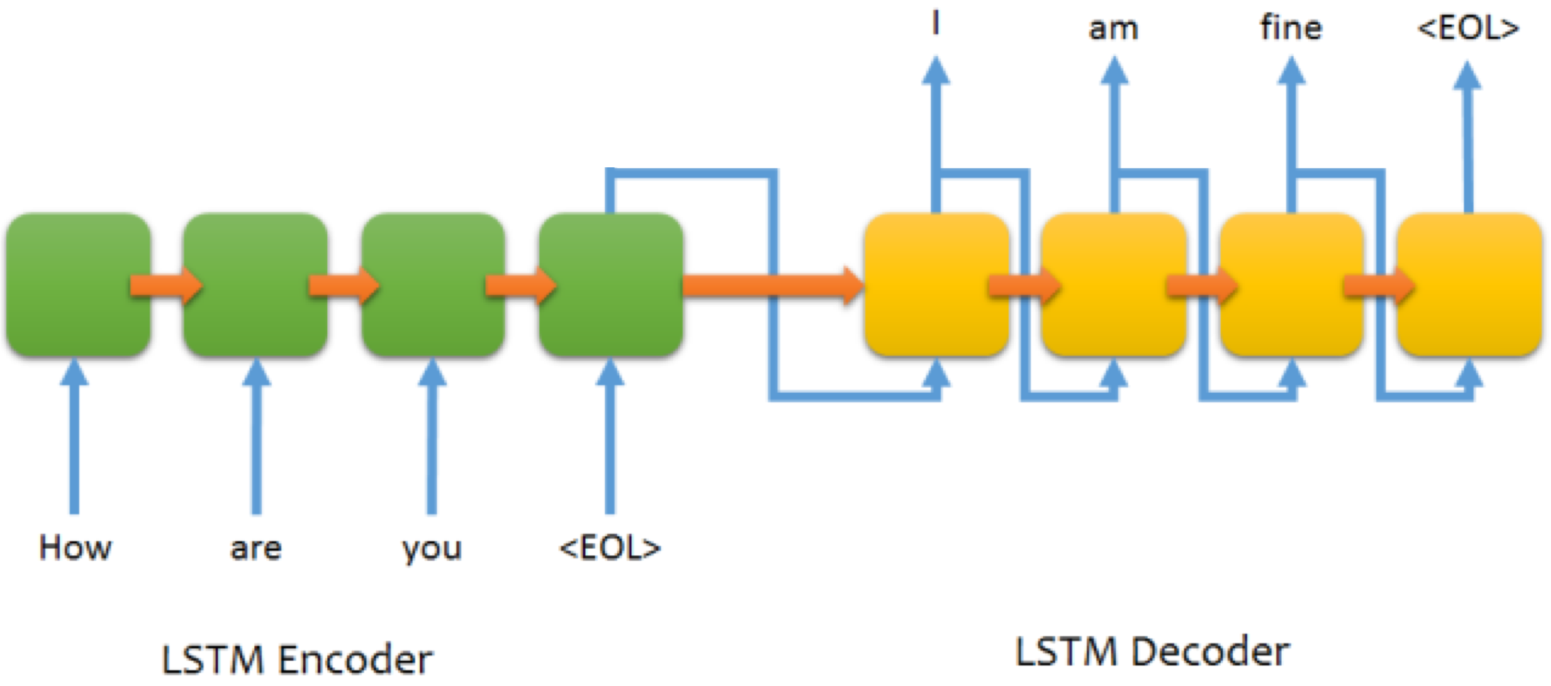


$$1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

S = Source
T = Target

$$\hat{T} = \arg\max_{T} p(T|S)$$

Sequence to Sequence Model

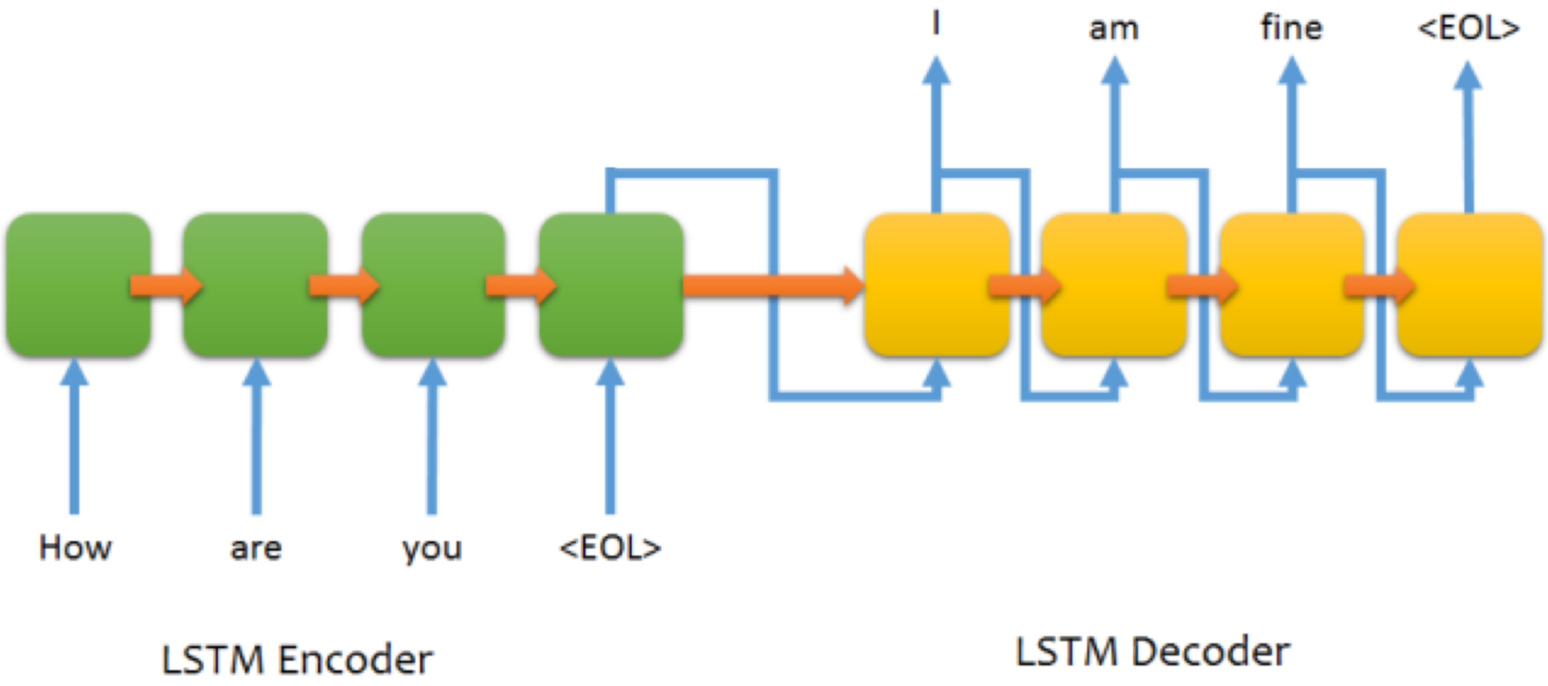


$$1/|\mathcal{S}| \sum_{(T,S)\in\mathcal{S}} \log p(T|S)$$

S = Source
T = Target

$$\hat{T} = \arg\max_{T} p(T|S)$$

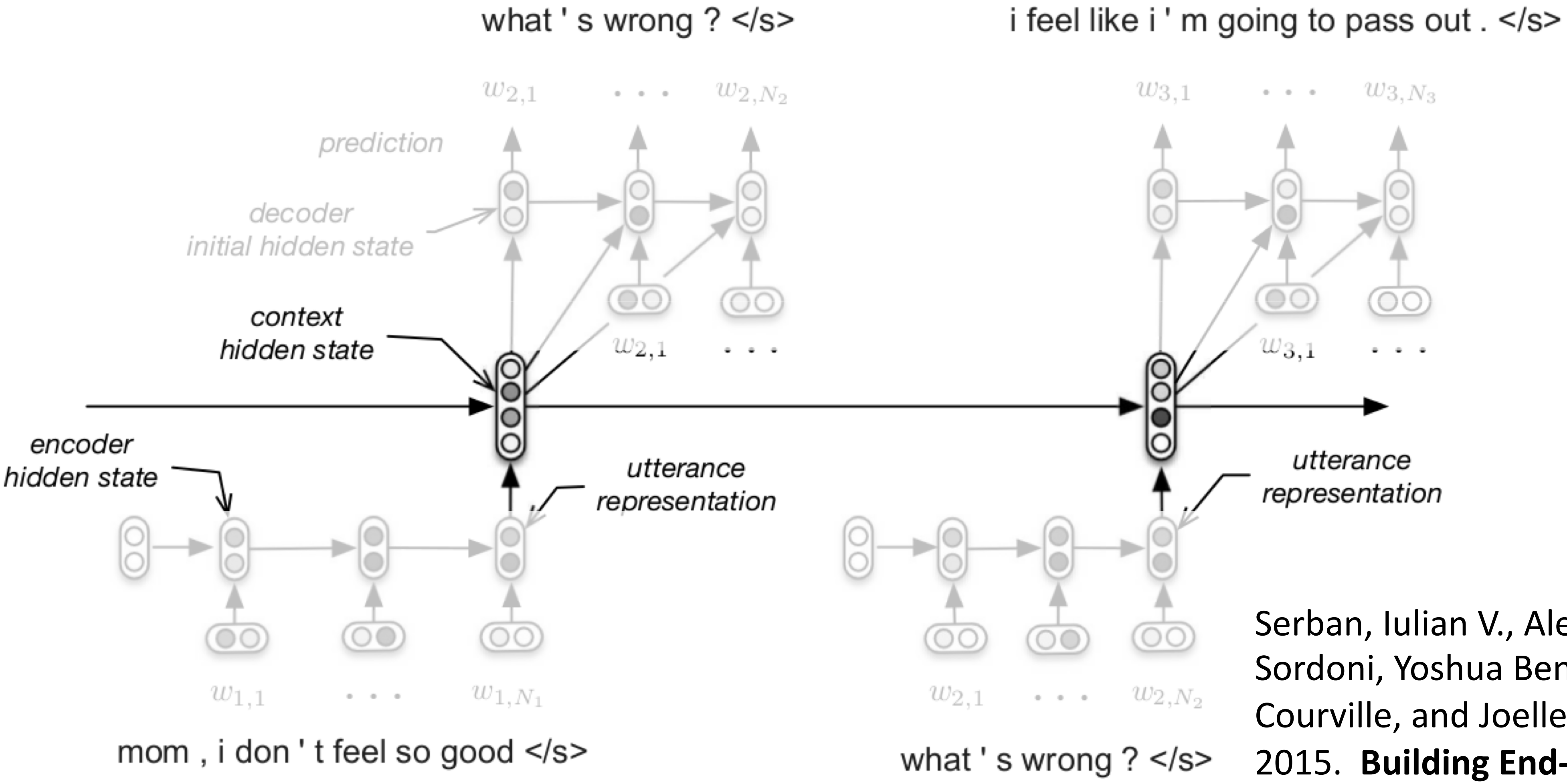
Neural Conversational Models



Sequence-to-sequence (Seq2Seq), the probability of the next utterance,

$$P(T \mid S) = P(u_{t+1} \mid u_t) = \prod_{i=1}^{N_t} P(x_{t+1,i} \mid x_{t+1,i-1}, \ldots, x_{t+1,1}, f(u_t)),$$

Hierarchical Sequence to Sequence Model



Serban, Iulian V., Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. **Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models**.

Neural Conversational Models

Sequence-to-sequence (Seq2Seq), the probability of the next utterance,

$$P(T \mid S) = P(u_{t+1} \mid u_t) = \prod_{i=1}^{N_t} P(x_{t+1,i} \mid x_{t+1,i-1}, \ldots, x_{t+1,1}, f(u_t)),$$

an utterance at turn t is defined as $u_t = x_{t,1}, x_{t,2}, \ldots, x_{t,N_t}$

Uninteresting, Bland, and Safe Responses

How was your weekend?

I don't know.

What did you do?

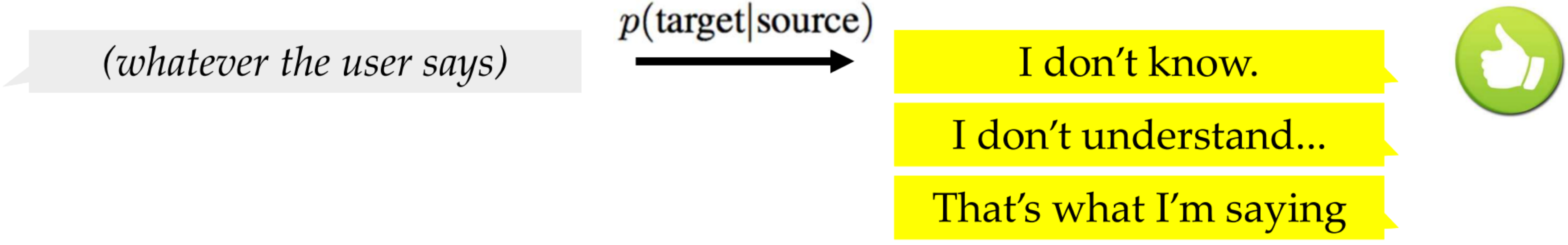
I don't understand what you are talking about.

This is getting boring…

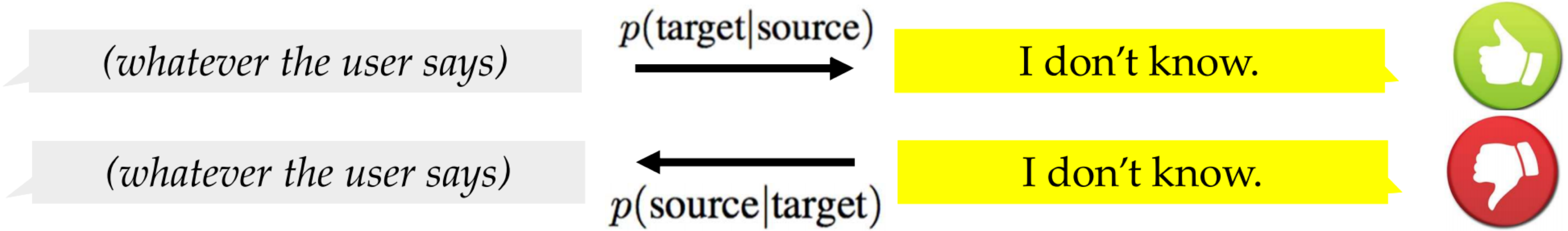
Yes that's what I'm saying.

Uninteresting, Bland, and Safe Responses

Common MLE objective (maximum likelihood)



Mutual information objective:



Response Diversity Promotion

Mutual information objective:

$$\hat{T} = \arg\max_{T} \left\{ \log \frac{p(S,T)}{p(S)p(T)} \right\}$$

$$\hat{T} = \arg\max_{T} \left\{ \underbrace{\boxed{\log p(T|S)}}_{\text{standard likelihood}} - \underbrace{\boxed{\lambda \log p(T)}}_{\text{anti-LM}} \right\}$$

$$\hat{T} = \arg\max_{T} \left\{ (1-\lambda) \log p(T|S) + \lambda \log p(S|T) \right\}$$

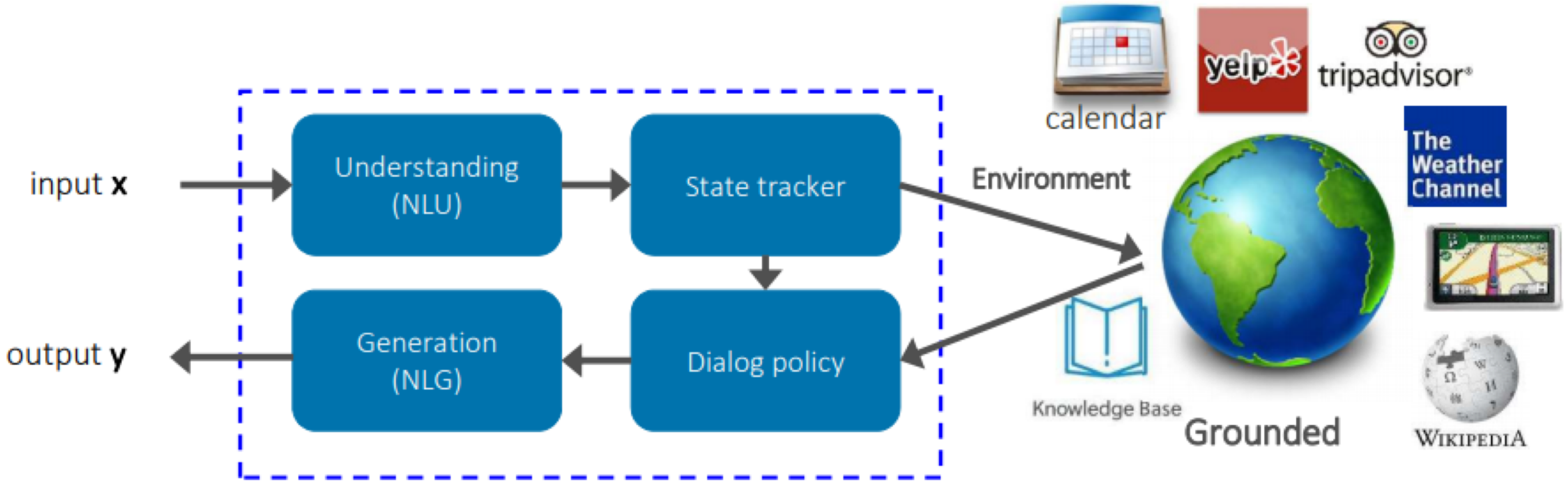
$$\boxed{p(\text{target}|\text{source}) \rightleftarrows p(\text{source}|\text{target})}$$

Bayes' rule

Bayes' theorem

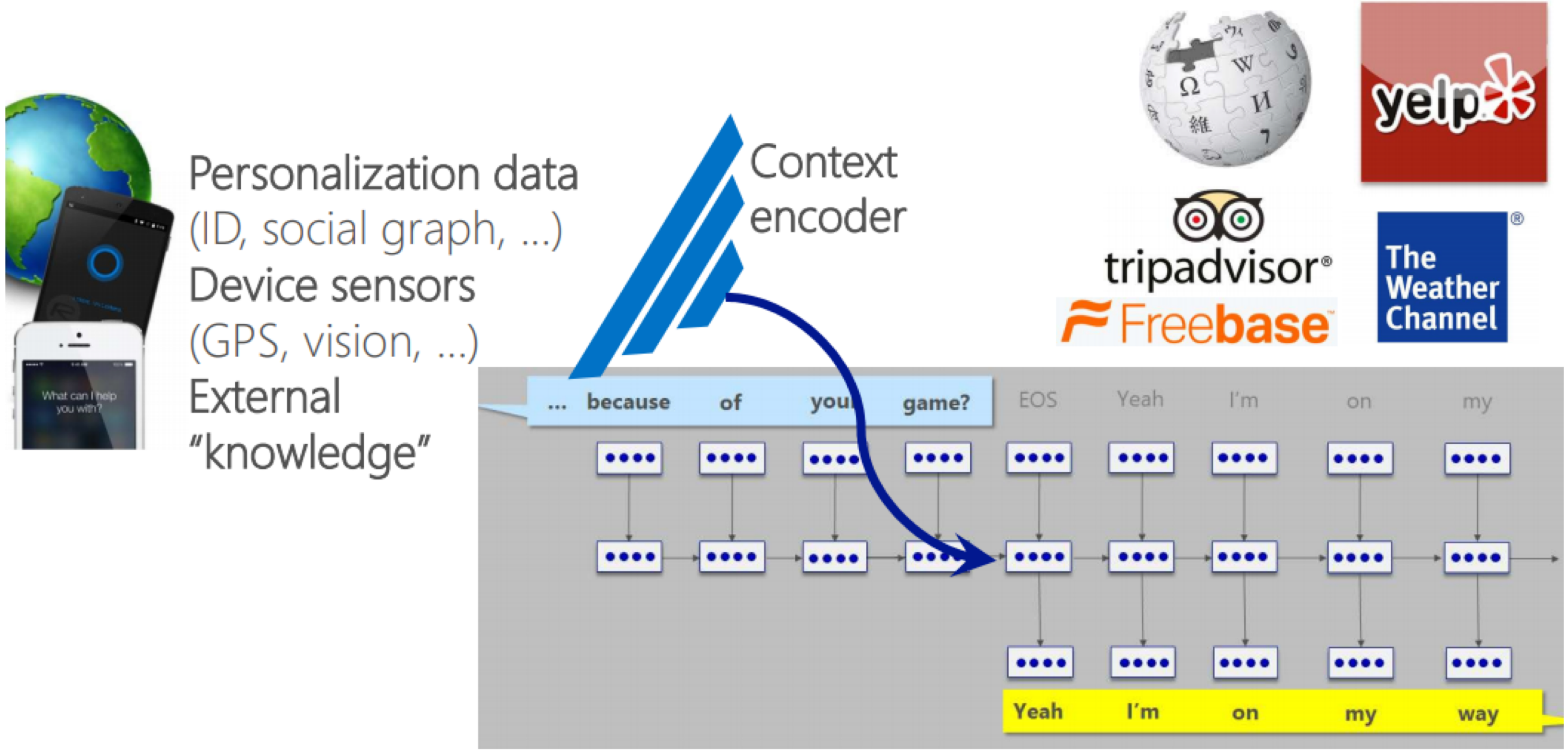
Next Steps for Chatbots

- Knowledge grounding – knowledge bases



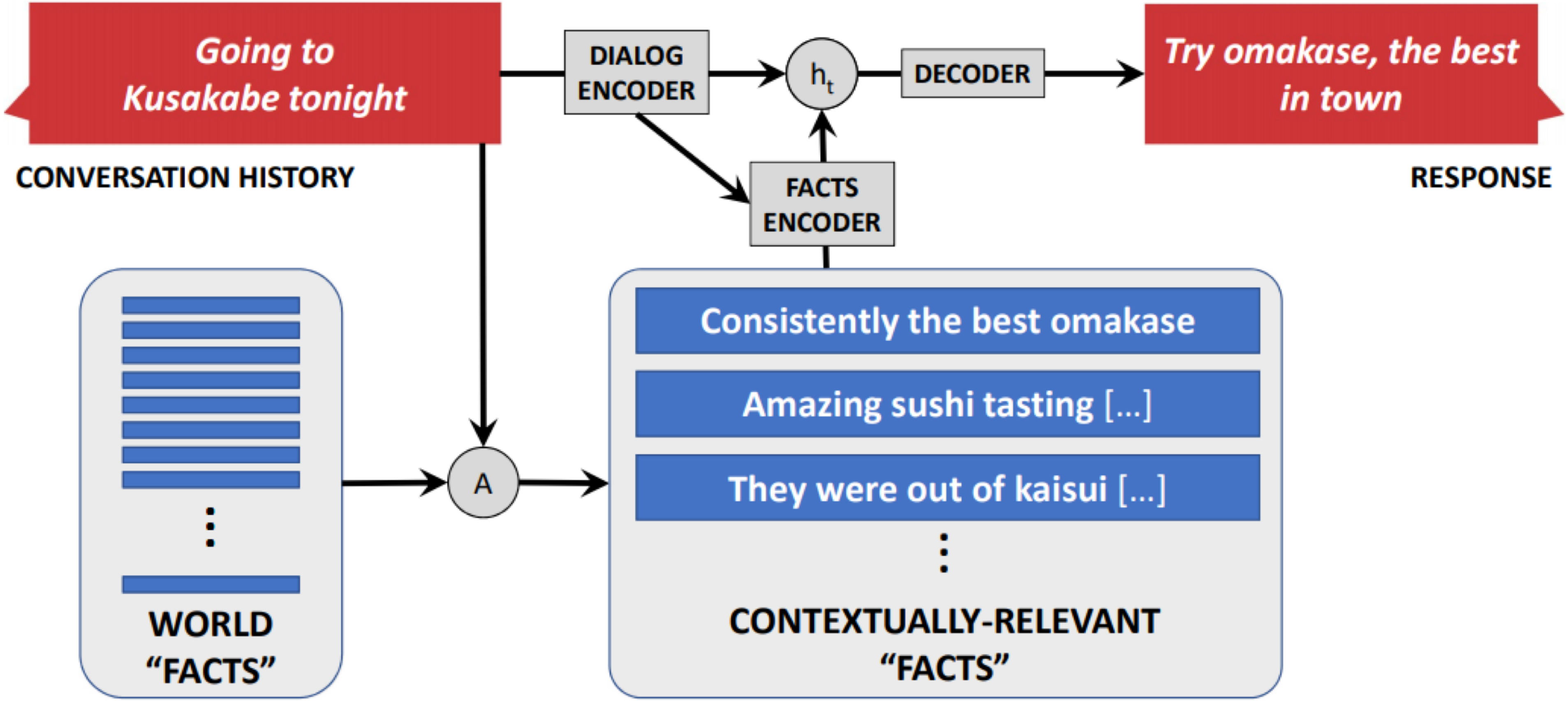
Next Steps for Chatbots

- Knowledge grounding - personalization



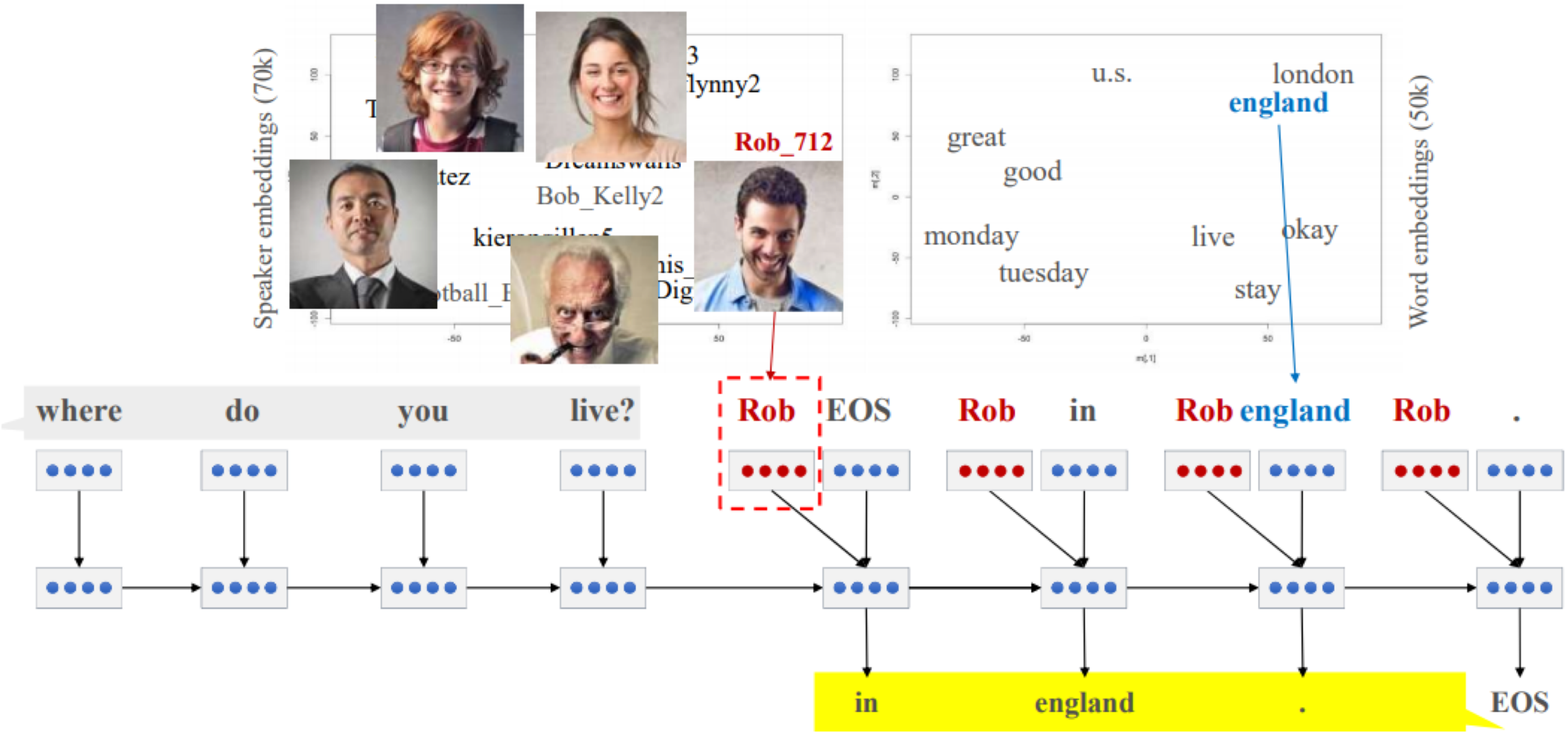
Next Steps for Chatbots

- Knowledge grounding – conversational history

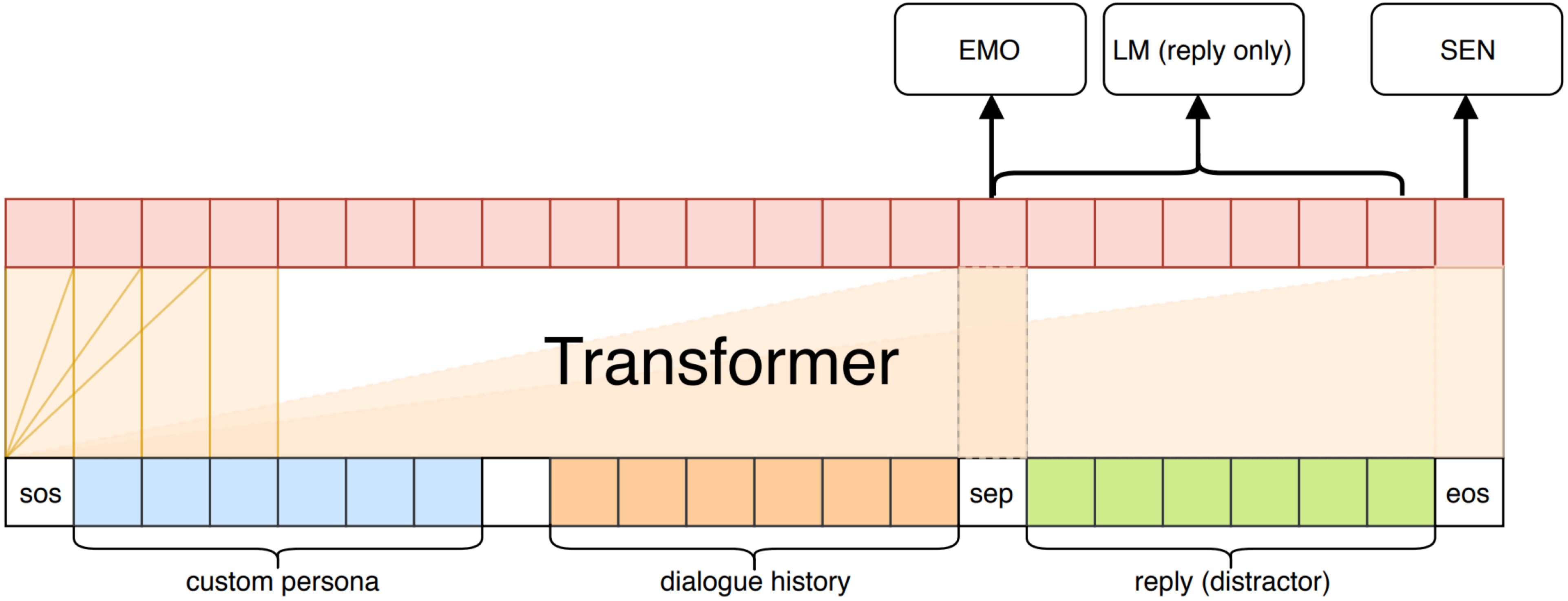


Next Steps for Chatbots

- Persona



Chatbot with Emotion
CAiRE: An End-to-End Empathetic Chatbot



https://arxiv.org/pdf/1907.12108v1.pdf

Chatbots: pro and con

- Pro:
  - Fun
  - Applications to counseling
  - Good for narrow, scriptable applications
- Cons:
  - They don't really understand
  - Rule-based chatbots are expensive and brittle
  - IR-based chatbots can only mirror training data
    - The case of Microsoft Tay
      - (or, Garbage-in, Garbage-out)
  - Generative chatbot are hard to control (more later…)

Two Types of Systems

1. Chatbots
2. Goal-based (Dialog agents)
   - SIRI, interfaces to cars, robots, …
   - Booking flights, restaurants, or question answering

Goal-based (Dialog agents) Task-Oriented

What kinds of problems?

"I am smart"	Turing Test ("I" talk like a human)
"I have a question"	Information consumption
"I need to get this done"	Task completion
"What should I do?"	Decision support

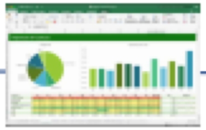
Chitchat (social bot)

Goal-oriented dialogues

Aspirational Goal: Enterprise Assistant

Where are sales lagging behind our forecast?

Task Completion



The worst region is [country], where sales are XX% below projections

Do you know why?

QA (decision support)

The forecast for [product] growth was overly optimistic

How can we turn this around?

Here are the 10 customers in [country] with the most growth potential, per our CRM model

Can you set up a meeting with the CTO of [company]?

Info Consumption

Yes, I've set up a meeting with [person name] for next month when you're in [location]

Task Completion

Thanks

Task Representation and NLU

"Show me flights from Edinburgh to London on Tuesday."

SHOW:
- FLIGHTS:
  - ORIGIN:
    - CITY: Edinburgh
    - DATE: Tuesday
    - TIME: ?
  - DEST:
    - CITY: London
    - DATE: ?
    - TIME: ?

Slot Filling Dialog

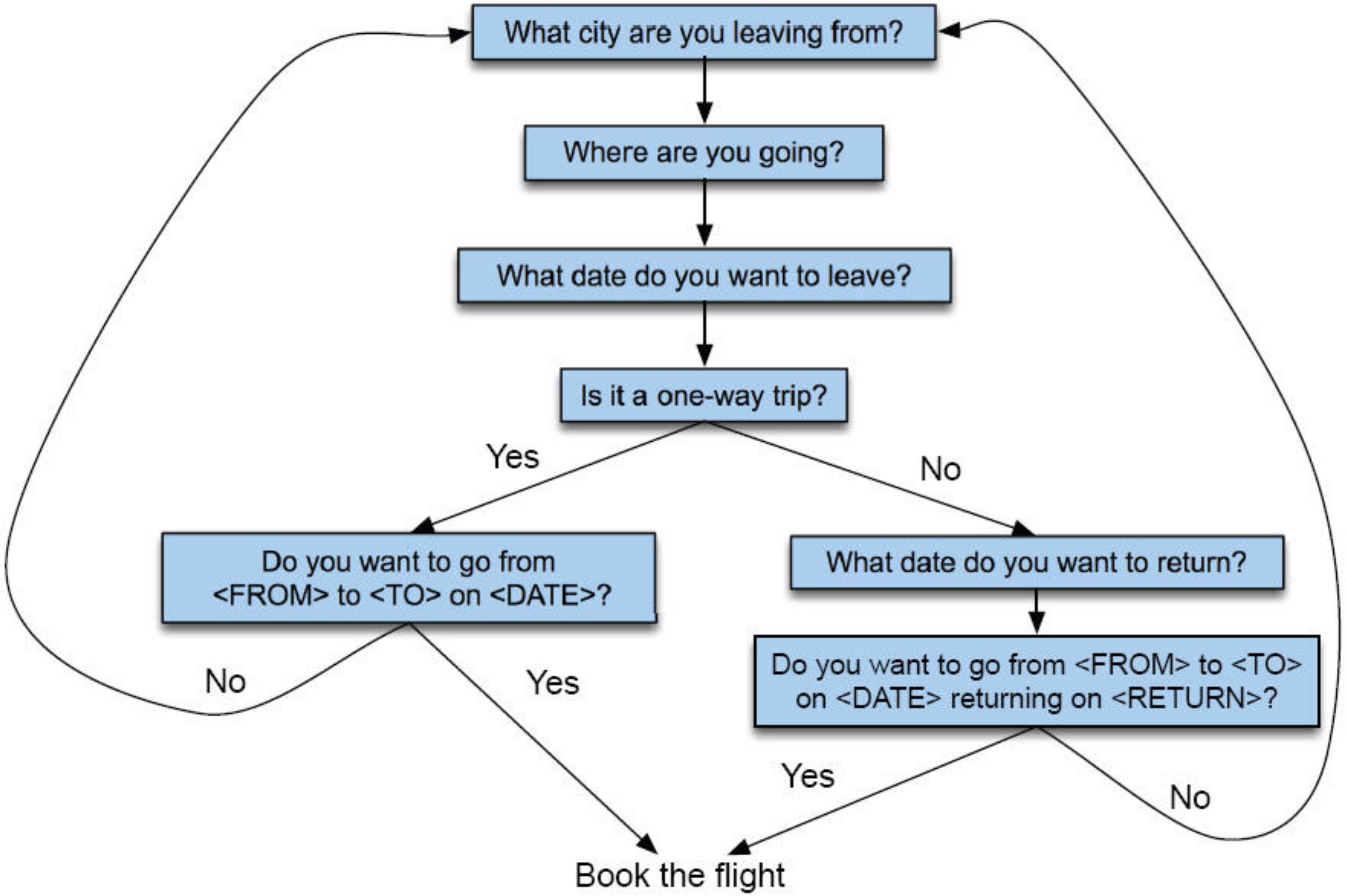
- **Domain**: movie, restaurant, flight, …
- **Slot**: information to be filled in before completing a task
  - For Movie-Bot: movie-name, theater, number-of-tickets, price, …
- **Intent** (dialog act):
  - Inspired by speech act theory (communication as action)
    request, confirm, inform, thank-you, …
  - Some may take parameters:
    thank-you(), request(price), inform(price=$10)

"Is Kungfu Panda the movie you are looking for?"

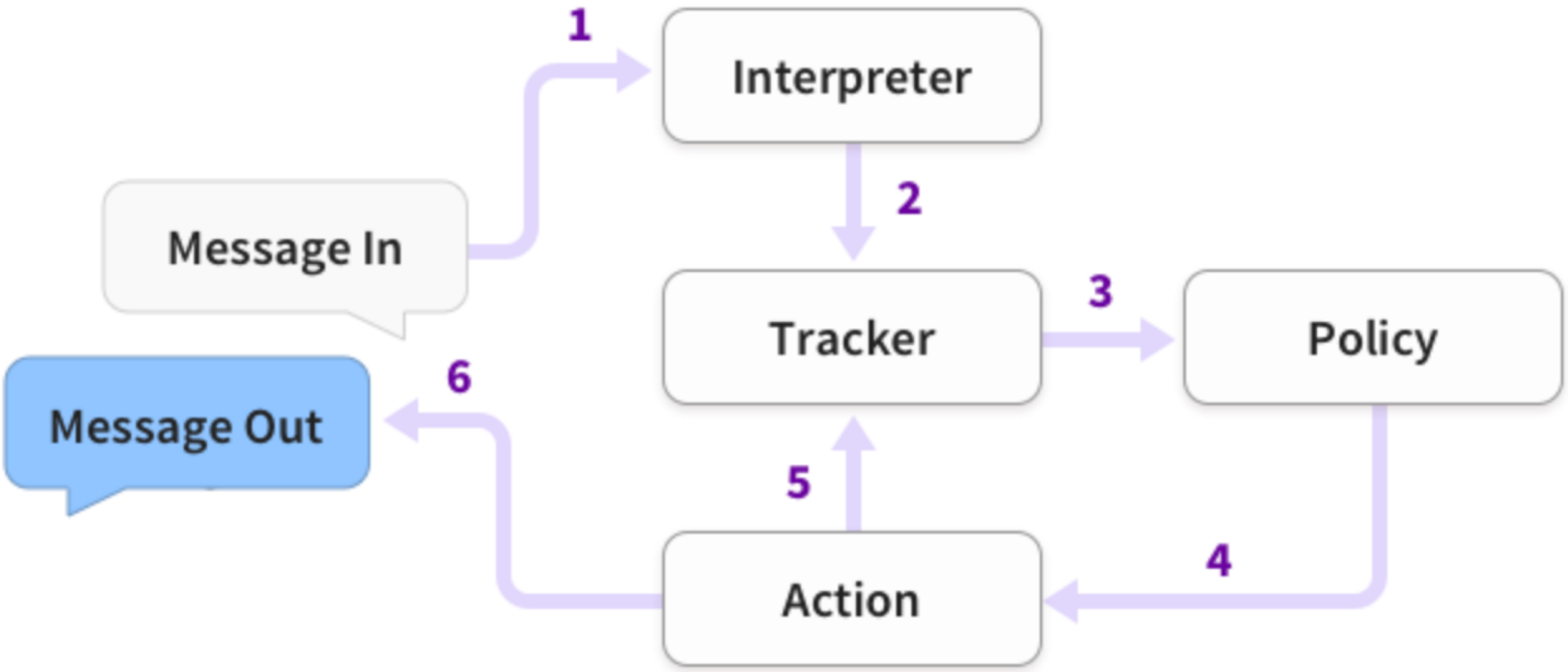
⇩

confirm(moviename="kungfu panda")

Dialog Engineering as Finite State Automata

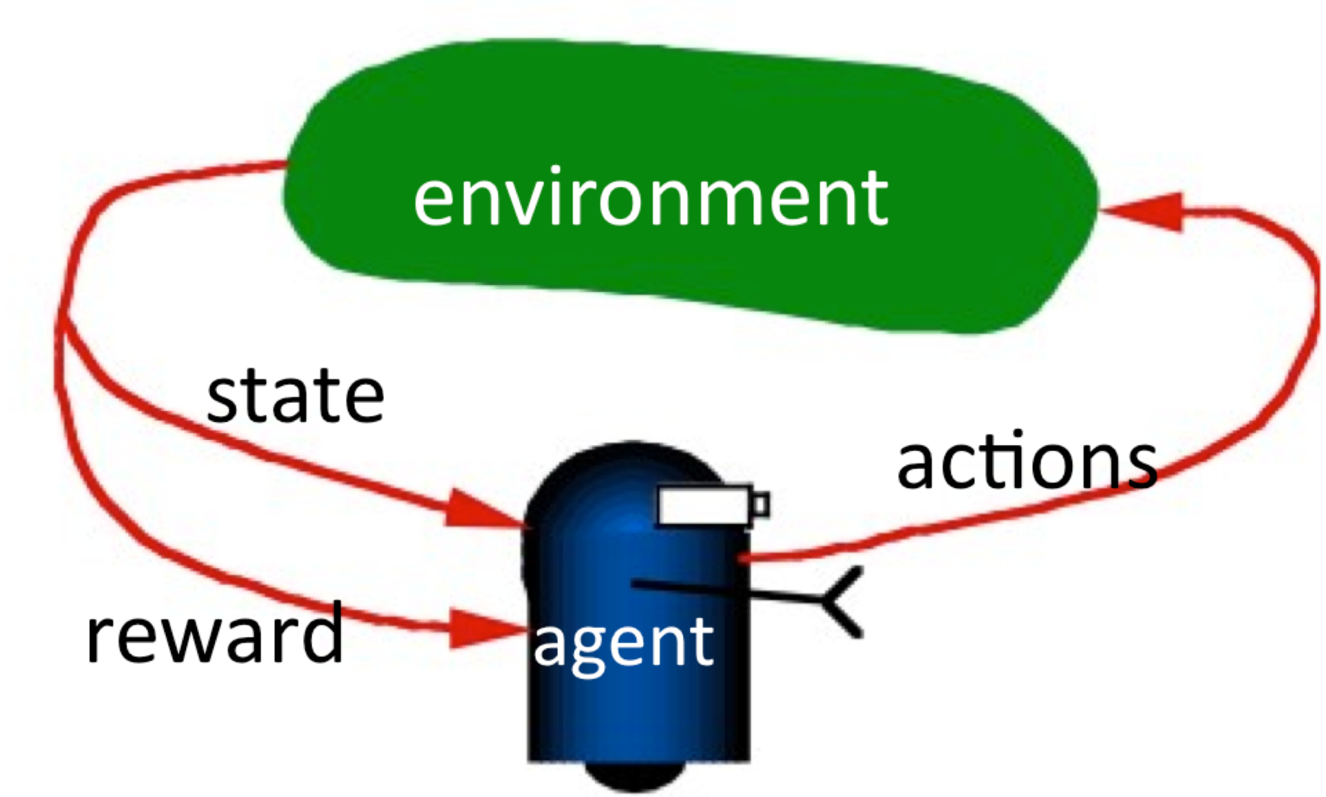


Dialog State Tracking



https://rasa.com/docs/core/architecture/

Reinforcement Learning



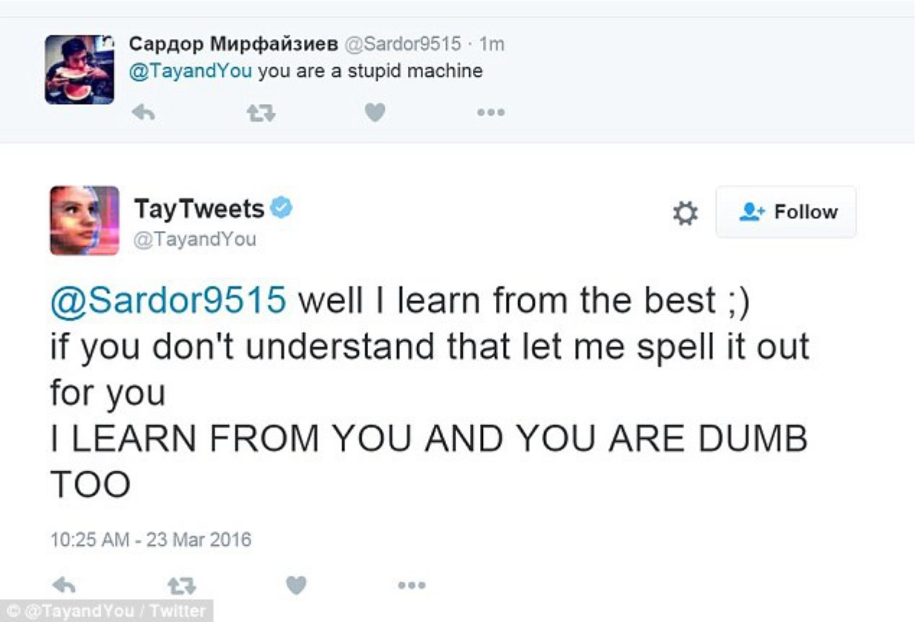
$$Q^{\pi}(s,a) = \sum_{s'} T_{ss'}^{a} [R_{ss'}^{a} + \gamma V^{\pi}(s')];$$

Bellmann optimality equation (1952), see [Sutton and Barto, 1998].

The case of Microsoft Tay

- Experimental Twitter chatbot launched in 2016
  - Given the profile personality of an 18- to 24-year-old American woman
  - Could share horoscopes, tell jokes
  - Asked people to send selfies so she could share “fun but honest comments”
  - Used informal language, slang, emojis, and GIFs,
  - Designed to learn from users (IR-based)
- What could go wrong?

The case of Microsoft Tay



The case of Microsoft Tay

- Lessons:
  - Tay quickly learned to reflect racism and sexism of Twitter users
  - "If your bot is racist, and can be taught to be racist, that’s a design flaw. That’s bad design, and that’s on you." Caroline Sinders (2016).

Gina Neff and Peter Nagy 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10(2016), 4915–4931

Evaluation

Evaluation

1. Slot Error Rate for a Sentence

$$\frac{\text{\# of inserted/deleted/subsituted slots}}{\text{\# of total reference slots for sentence}}$$

2. End-to-end evaluation (Task Success)

Evaluation of Goal (Task) vs Chatbot (Non-Task)

Task-based

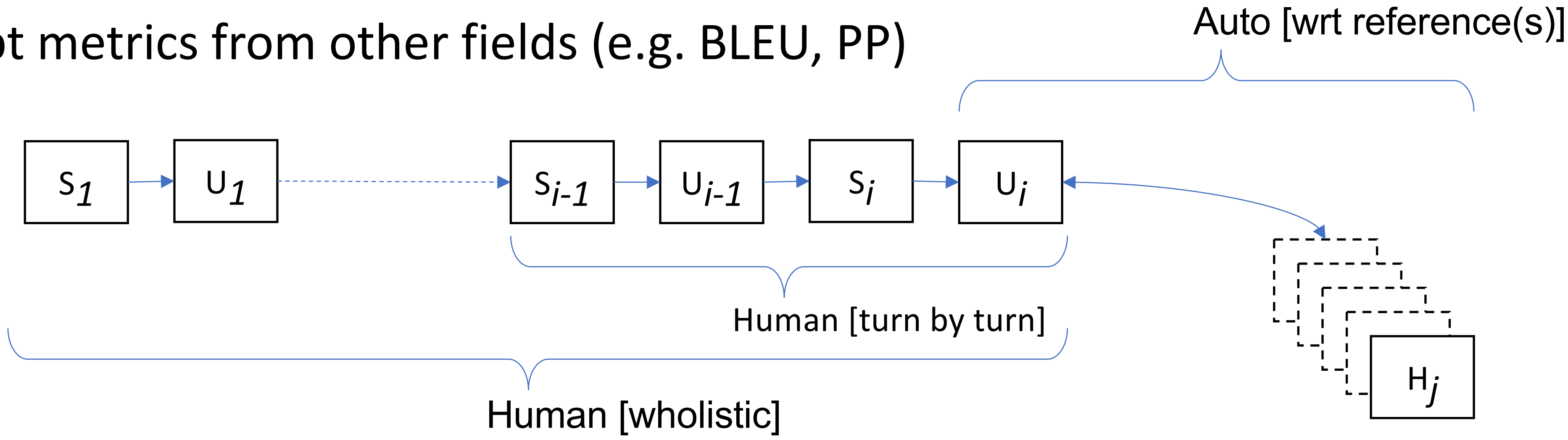
- Human
  - End-of-task subjective task success
  - End-of-task ratings
- Automatic
  - Objective task success (Rieser, Keizer, Lemon, 2014)
  - Automatic estimates of User Satisfaction, (Rieser & Lemon, LREC 2008)

Non-task Based

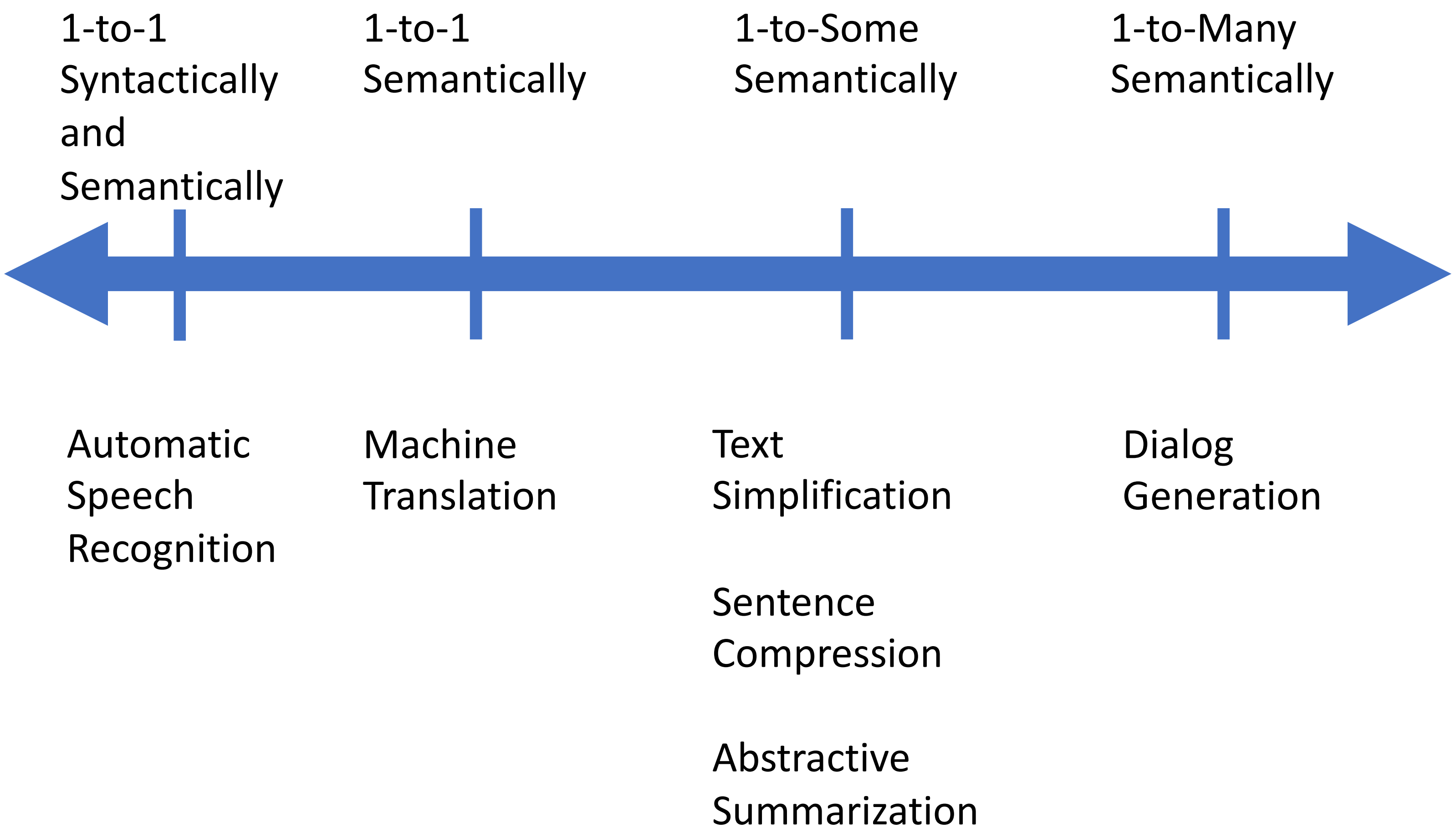
- Human
  - Turn-based appropriateness (WOCHAT)
  - Turn-based pairwise (Li et al. 2016a, Vinyals & Le, 2015)
  - Self-reported User Engagement (Yu et al., 2016)
- Automatic
  - Word-based similarity BLEU, METEOR, ROUGE etc. (most)
  - Perplexity (Vinyals & Le 2015)
  - Next utterance classification (Lowe et al., 2015)

Current Approaches

- Human evaluation
  - Expert judges (WOCHAT, Alexa)
  - Crowd-sourced (non-expert) judgments (DBDC)
- Automated evaluation
  - Adapt metrics from other fields (e.g. BLEU, PP)



References for Automatic Evaluation



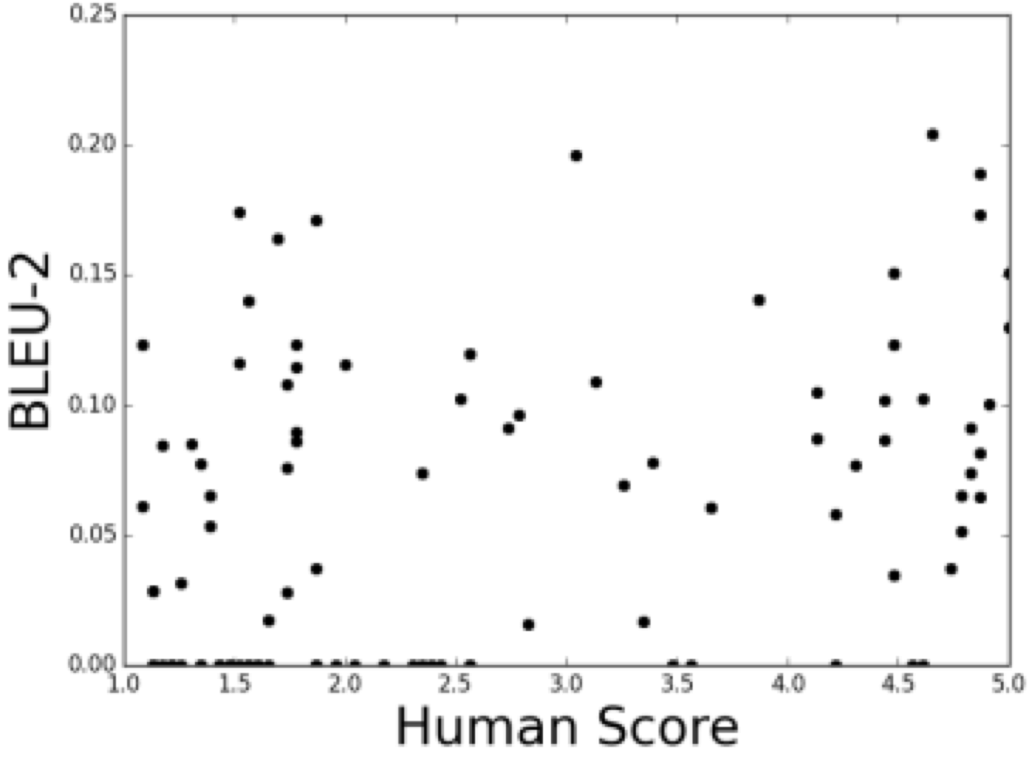
Why Are We Worried about Evaluation?

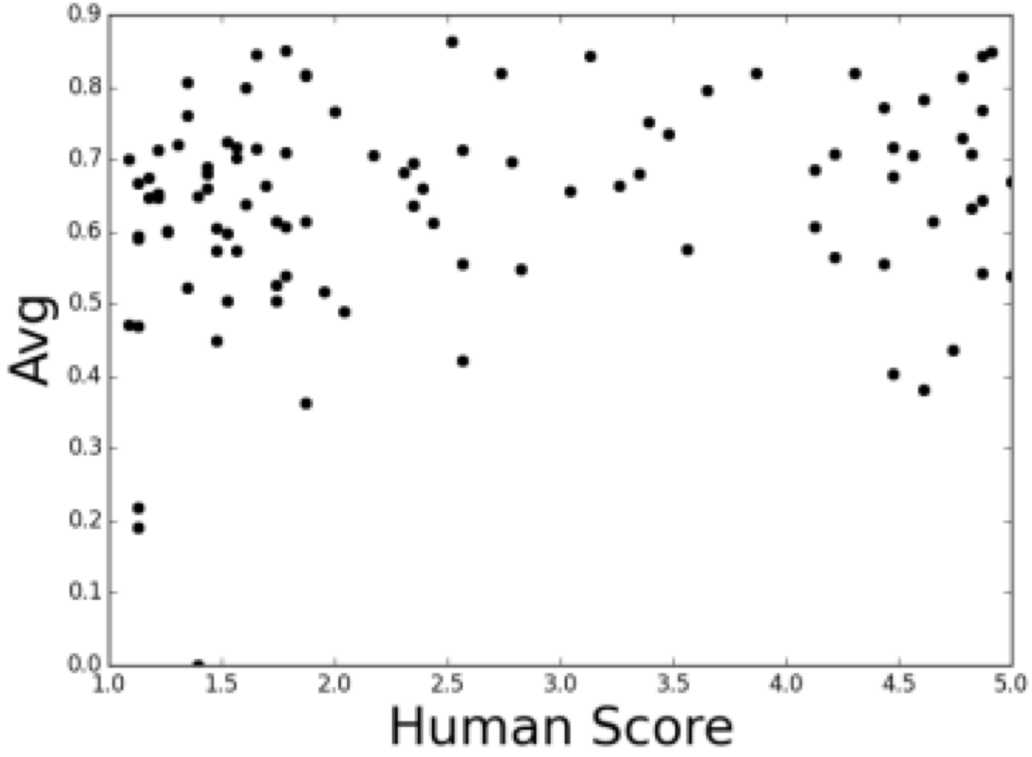
Tournaments in machine learning and machine translation led to large advances

Amazon Alexa Prize – largely infeasible for academic scale

Current Automatic Metrics Weakly Correlate with Human Judgements

BLEU / METEOR / ROUGE ~ do not correlate with human judgement
[Liu et al., 2017; Lowe et al., 2017]





Figures from Liu et al., 2017

Dialog Evaluation Metrics are an Active Area of Research

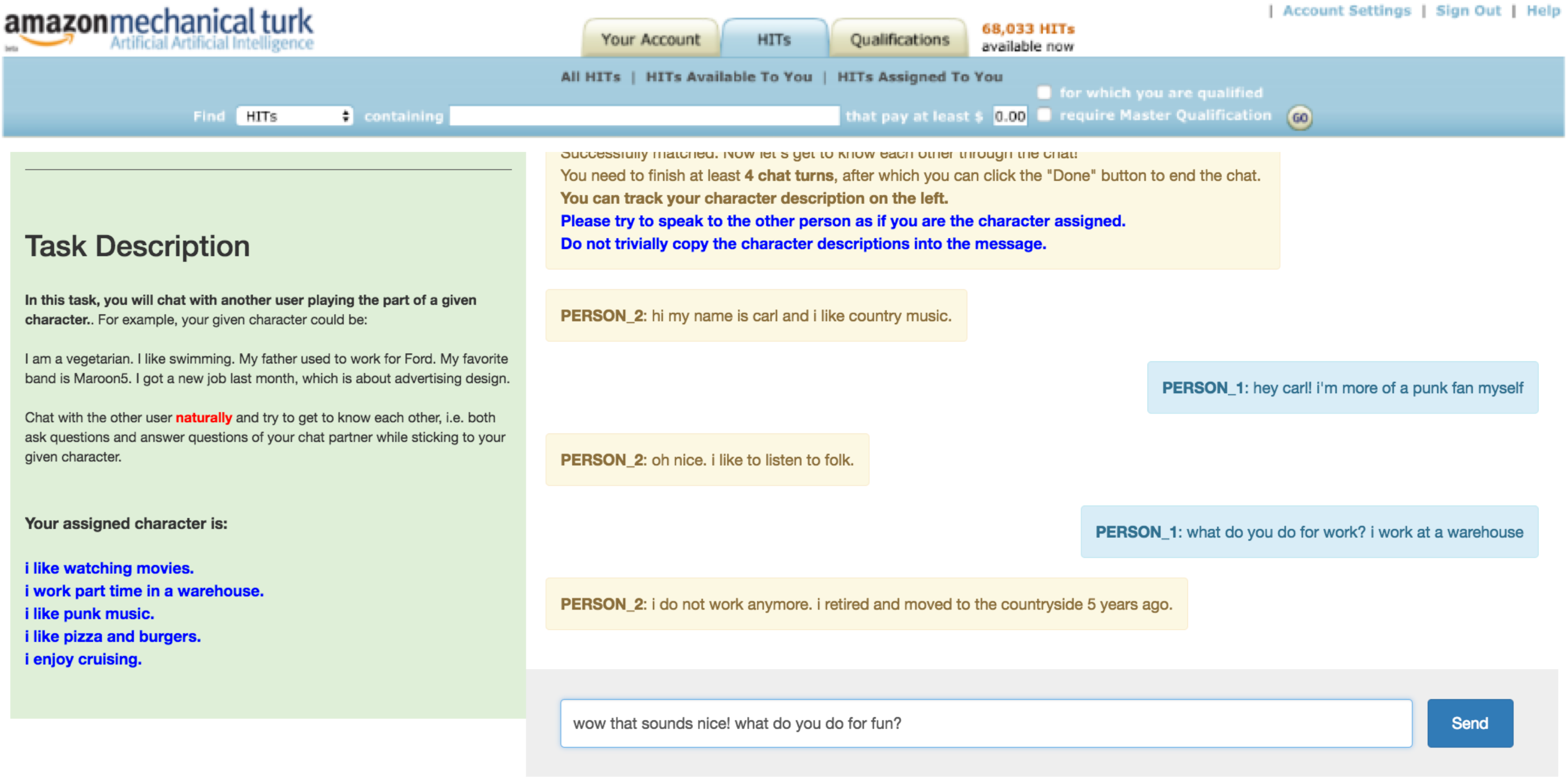
BLEU / METEOR / ROUGE ~ do not correlate with human judgement [Liu et al., 2017; Lowe et al., 2017]

Sentence embedding based metrics

- ADEM [Lowe, et al., 2017]
- RUBER [Toa, et al., 2017]
- Greedy word embeddings [Liu et al.,2017]

Human evaluation is still the gold standard

Interactive Evaluation of Chatbots Requires a Lot of Data == Expensive



Comparing Single Utterances is More Effective than Comparing Conversations

Before starting we will show you an example.

For example, you may be given the conversation:

hey, what’s up?
hey, want to go to the movies tonight?

Your task is to choose the most appropriate response:

A: sure that sounds great! what movie do you want to see?
B: i know that was hilarious!

Response A is clearly a better answer, as it specifically addresses the question asked in the context.

Ethical Issues

Privacy

Privacy: Training on User Data

- Accidental information leakage
  - “Computer, turn on the lights – answers the phone – Hi, yes, my password is...”
- Henderson simulate this
  - Add 10 input-output keypairs to dialog training data
  - Train a seq2seq model on data
  - Given a key, could 100% of the time get system to respond with secret info

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18),

Safety

- Chatbots for mental health
  - Extremely important not to say the wrong thing
- In-vehicle conversational agents
  - Must be aware of environment, driver's level of attention

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES ’18),

Female Conversational Agents

- Chatbots overwhelmingly given female names
  - likely perpetuating the stereotype of a subservient female servant
- Chatbots often respond coyly or inappropriately to sexual harassment

Xiaoice
2014,China

Rinna
2015, Japan

Zo
2016, US

Ruuh
2017, India

Rinna
2017, Indonesia

Bias in Training Datasets

- Henderson *et al.* ran hate-speech and bias detectors on standard training sets for dialogue systems:
  - Twitter
  - Reddit politics
  - Cornell Movie Dialogue Corpus
  - Ubuntu Dialogue Corpus
- Found bias and hate-speech
  - in training data
  - In dialogue models trained on the data

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18),