Generation based Conversational AI

An Introduction

Thy Thy Tran, PhD

06.2021

Outline

- Introduction
- Building a Conversational Agent
- Summary

Two Main Types of Dialogue Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

How is the weather today?

Right now is 21 degrees with cloudy.

Is it good to go out for dinner at 6 pm?

Absolutely, the sky will be clear by 6.

Ok. Then please book me a table for 2 at Sushi Bar at 6 pm

Confirm a table for 2 at 6 pm. Anything else?

That's it. Thanks.

Happy to help

Information consumption

Decision Support

Task completion

[Fung et al., 2020]

Task-oriented conversations

help users achieve some goals/tasks

complete requests with **least turns**

Open-domain chit chat

free-form, open-domain

engage users for long conversations

Conversational AI Applications

How is the weather today?

Right now is 21 degrees with cloudy.

Is it good to go out for dinner at 6 pm?

Absolutely, the sky will be clear by 6.

Ok. Then please book me a table for 2 at Sushi Bar at 6 pm

Confirm a table for 2 at 6 pm. Anything else?

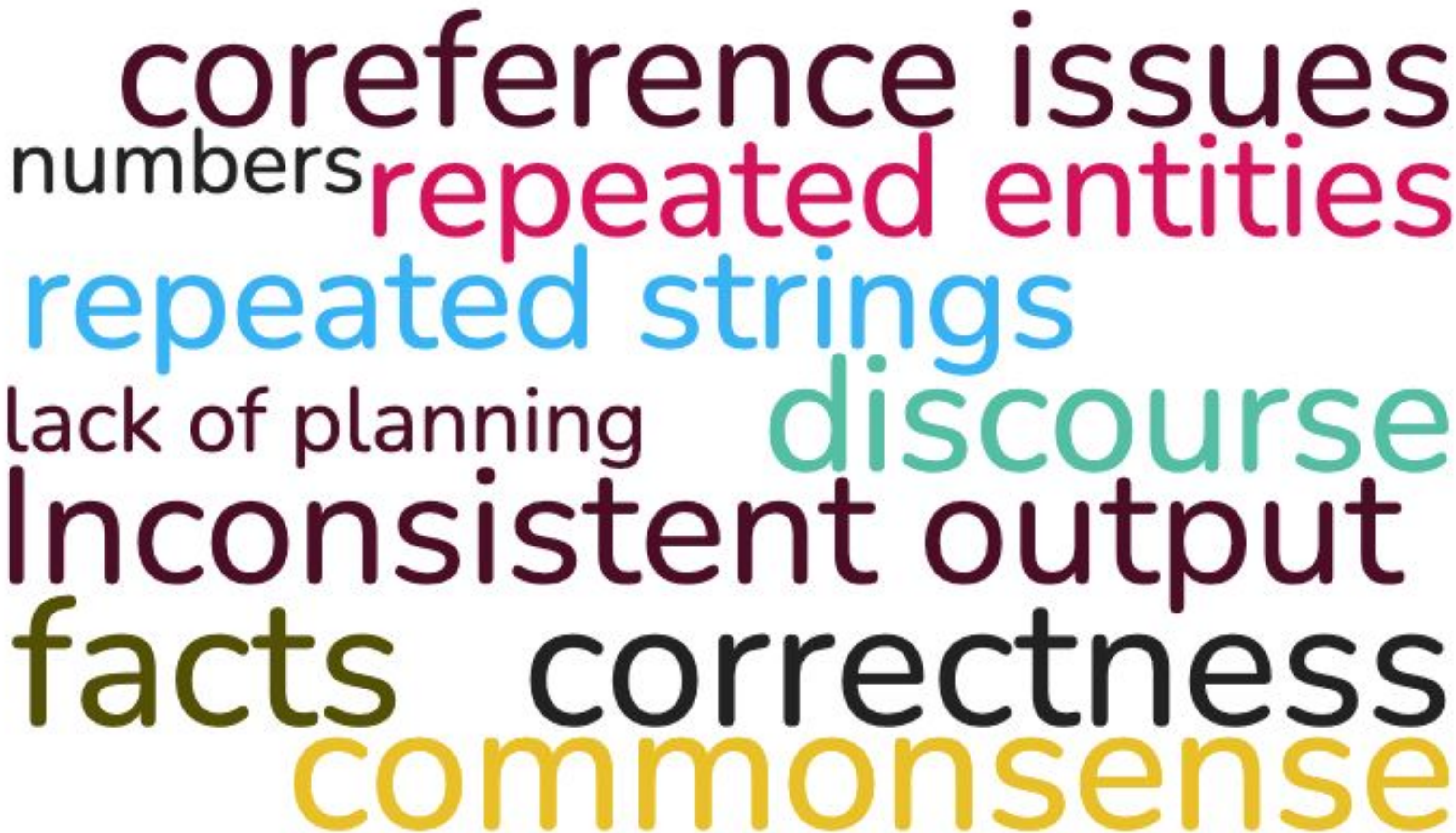
That’s it. Thanks.

Happy to help

- Lots of applications
  - Personal assistants [Siri, Google, Alexa, Cortana, etc.]
  - Education [Kerry et al., 2008, Mesgar et al., 2019]
  - Health care: therapy chatbot [Fitzpatrick et al., 2017]
  - Customer service: [Cui et al., 2017]
  - ...

Challenges of Generation Dialogue Systems

TECHNISCHE UNIVERSITÄT DARMSTADT



Building a Conversational Agent

4-step recipe
[Fung et al., 2020]

1. Data
2. Model
3. Training
4. Evaluation

Building a Conversational Agent

TECHNISCHE UNIVERSITÄT DARMSTADT

4-step recipe
[Fung et al., 2020]

1. Data
2. Model
3. Training
4. Evaluation

Types of Conversational Corpora

- Machine to machine
  - Generated from dialog templates
  - Issues: data quality, naturalness, noises
- Human to machine (MultiWOZ, DSTC)
  - Collected from existing dialog systems
  - Issues: limited domains, biases, noises
- Human to human (Twitter, Reddit, OpenSubtitles, Persona Chat,)
  - Twitter, Reddit, etc.
  - Customer service
  - Issues: small size, limited domains

Some Terms

TECHNISCHE UNIVERSITÄT DARMSTADT

4-step recipe
[Fung et al., 2020]

1. Data
2. Model
3. Training
4. Evaluation

Dialogue history

How is the weather today?

Right now is 21 degrees with cloudy.

Is it good to go out for dinner at 6 pm?

Absolutely, the sky will be clear by 6.

Ok. Then please book me a table for 2 at Sushi Bar at 6 pm

Confirm a table for 2 at 6 pm. Anything else?

That's it. Thanks.

Happy to help

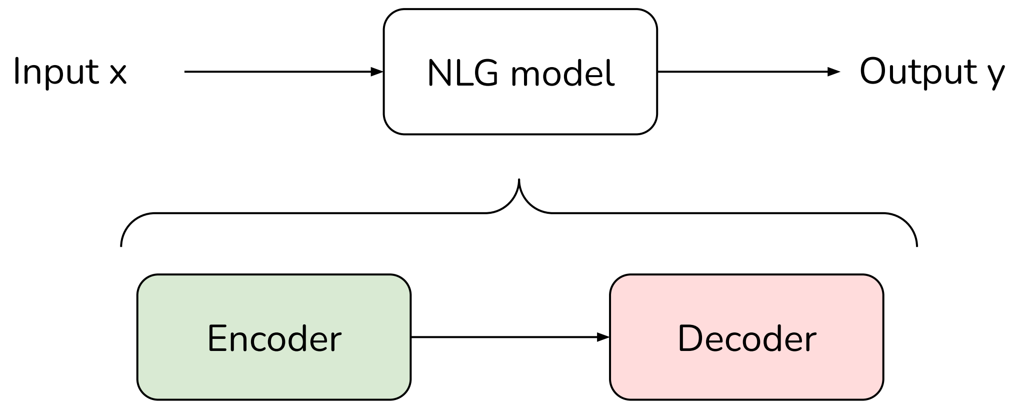
Utterance

Response

Basic Architecture of Generation based Conversational Models

4-step recipe

1. Data
2. Model
3. Training
4. Evaluation

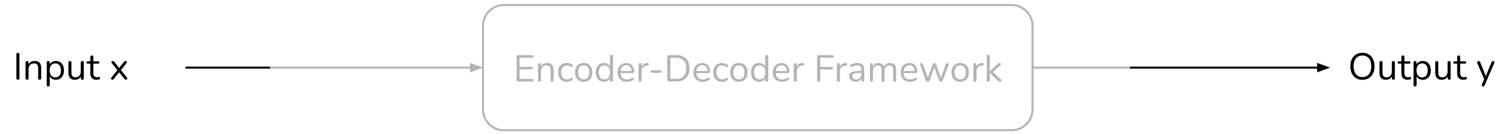


Understanding input (natural language understanding)

Generating the output based on

- the input
- target task

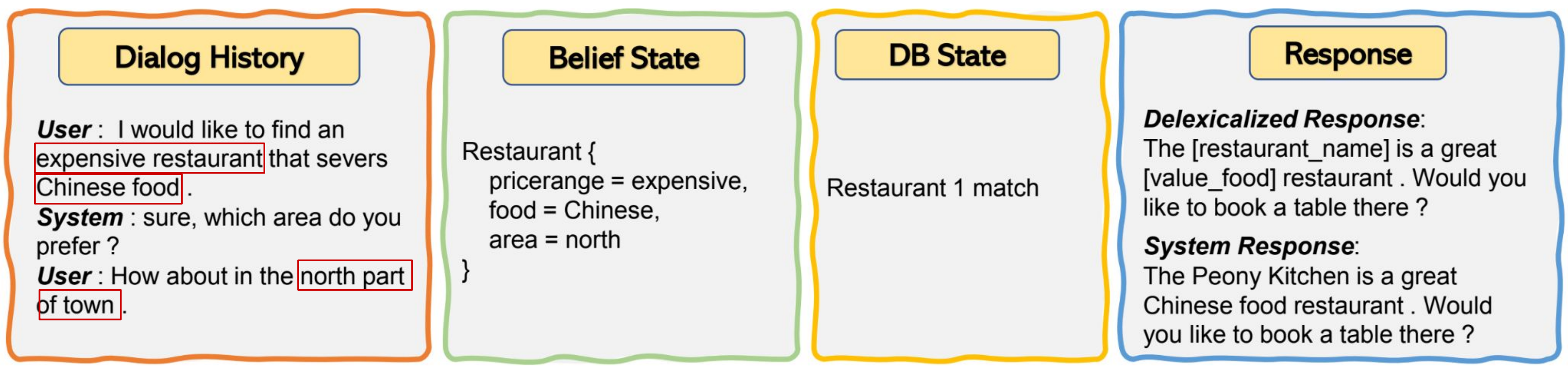
Input and Output



- Open-domain chit chat
  - Input
    - Dialogue history
  - Output
    - Response
- Task-oriented dialogs
  - Input
    - Dialogue history
    - Belief state
    - Database/API results (state)
  - Output
    - Response
    - Belief state
    - Database query
    - API Service

Input and Output: Task Oriented Dialogue Systems

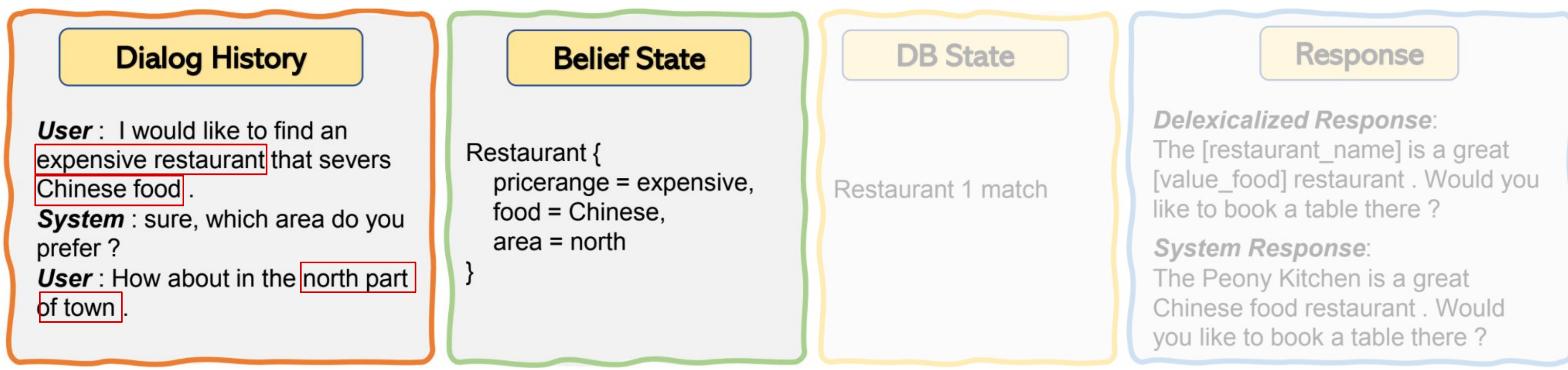
Each input/output is by itself a sequence of tokens



[Peng et al., 2020]

Input and Output: Task Oriented Dialogue Systems

Each input/output is by itself a sequence of tokens

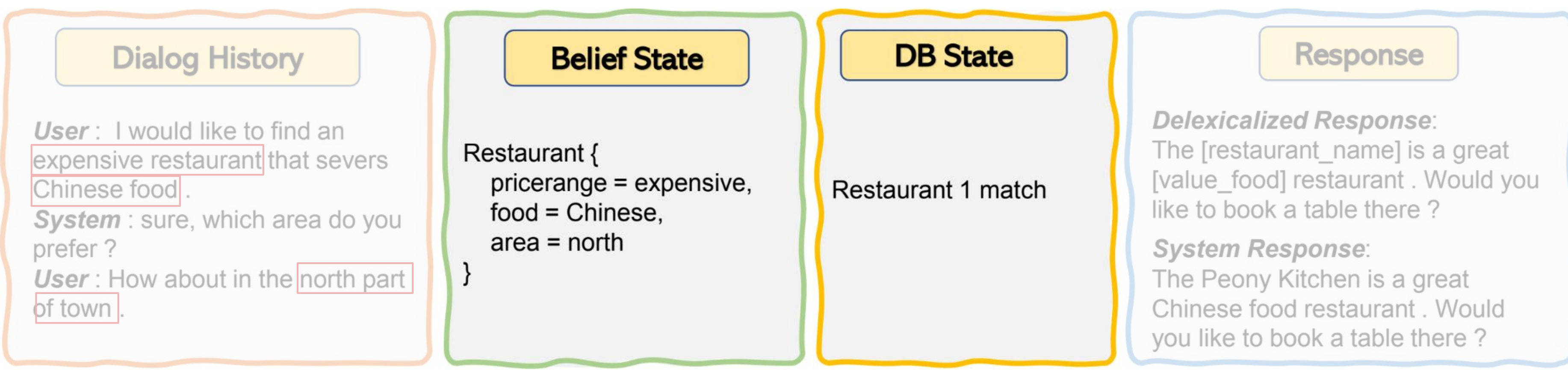


1

[Peng et al., 2020]

Input and Output: Task Oriented Dialogue Systems

Each input/output is by itself a sequence of tokens



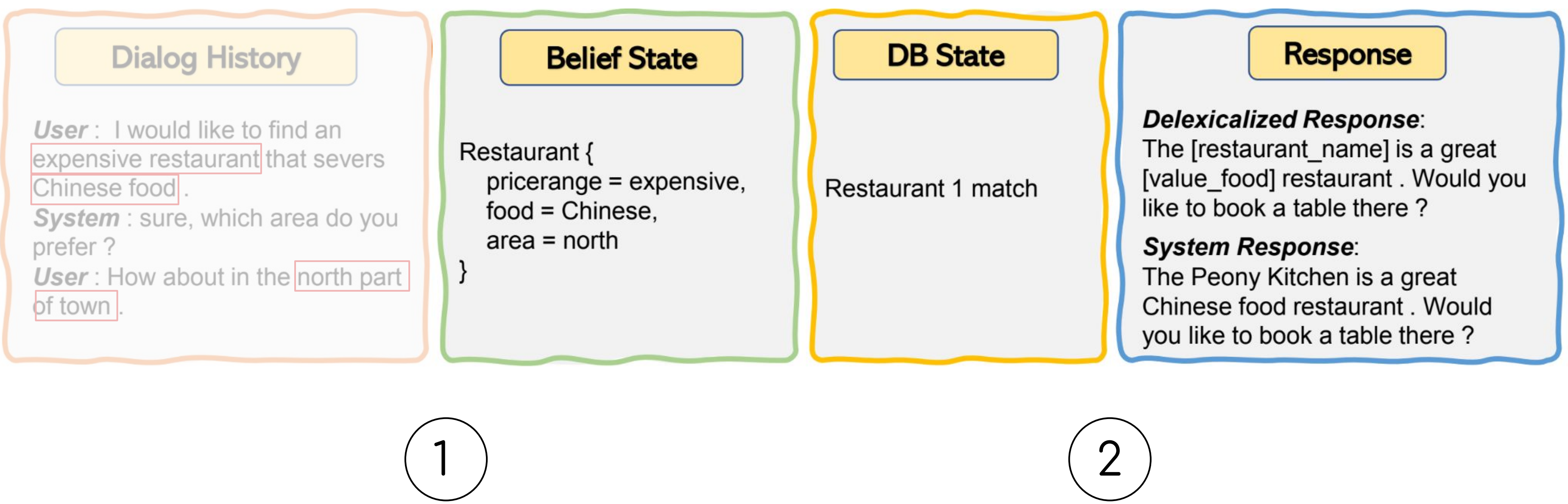
1

- 0 match
- >1 match

[Peng et al., 2020]

Input and Output: Task Oriented Dialogue Systems

Each input/output is by itself a sequence of tokens

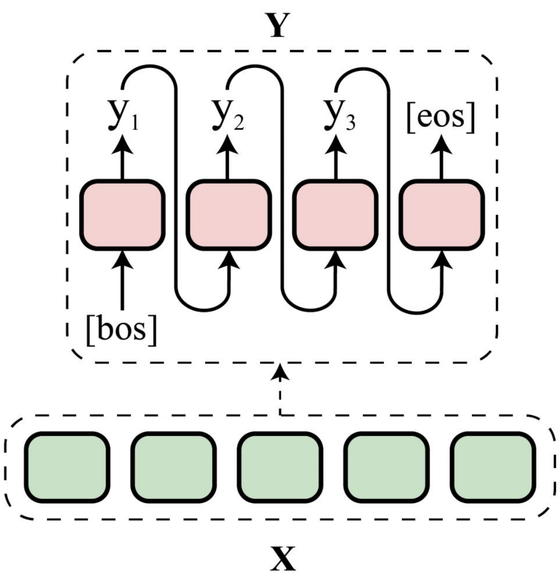


2.1 system action

[Peng et al., 2020]

Encoder-Decoder

Input x → Encoder-Decoder Framework → Output y



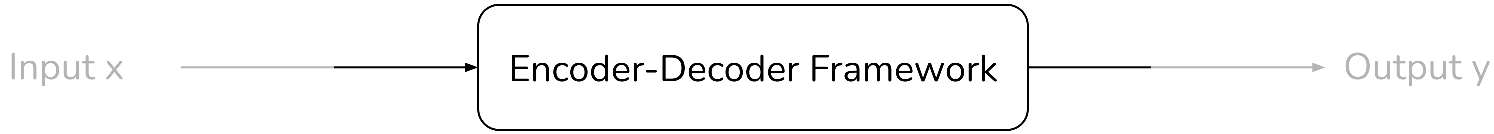
Autoregressive Generation [AG]

Issue with AG

- Error accumulation
  - worse generation at one step → even worse at following steps

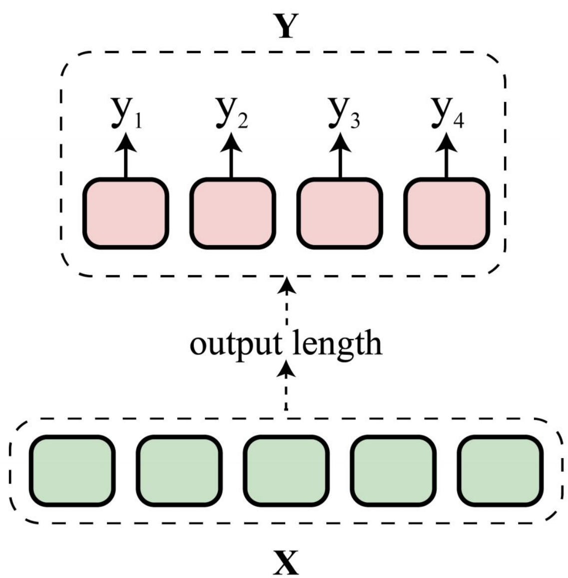
[Su et al., 2021]

Encoder-Decoder



Issues with NAG

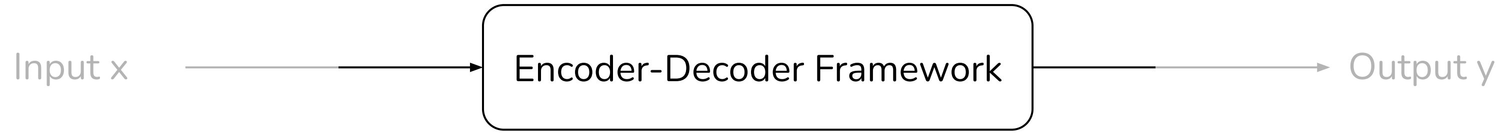
- Need to know target sequence length to generate all words in parallel
- Token repetition: conditional independence
  → repeat high probability tokens

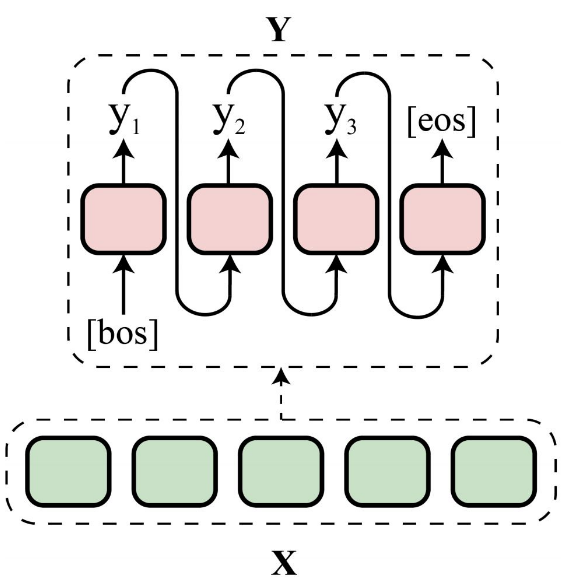


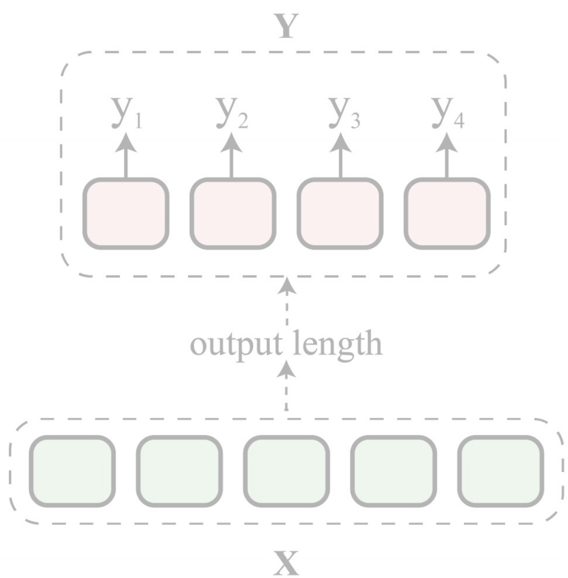
Non-Autoregressive Generation [NAG]

[Su et al., 2021]

Encoder-Decoder







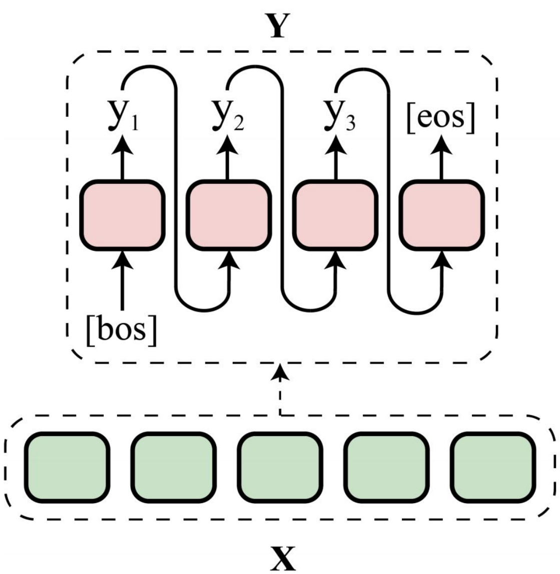
Autoregressive Generation [AG]

Non-Autoregressive Generation [NAG]

[Su et al., 2021]

Output Generation

Input x → Encoder-Decoder Framework → Output y



Autoregressive Generation [AG]

[Su et al., 2021]

- Starts with [bos] (begin-of-sequence)
- At each step
  - take previous generated token
  - generate a distribution over the vocabulary
- Stops by [eos] (end-of-sequence)
  - Terminate when [eos] is predicted
  - Stop when max target sequence length is reached

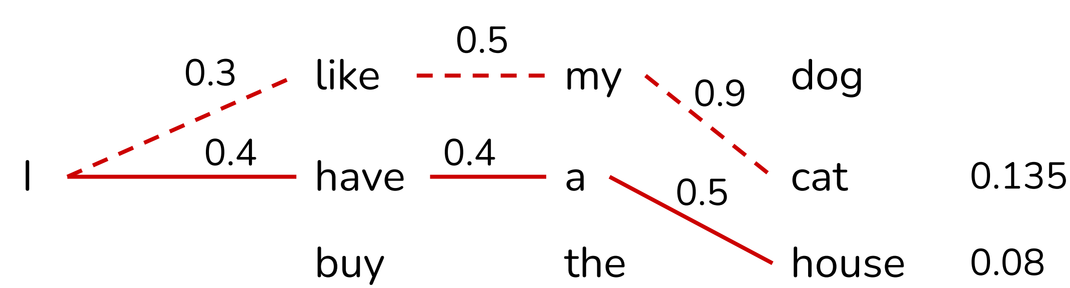
Output Generation: Decoding Strategies

	Greedy	Beam Search	Top-k Sampling	Top-p (Nucleus) Sampling
At each step	Pick the best word	Try a few best words	Random sample from top-k	smallest set with cumulative probability > p
Output	One sequence	Several partial sequences	One sequence	One sequence

Output Generation: Greedy Decoding

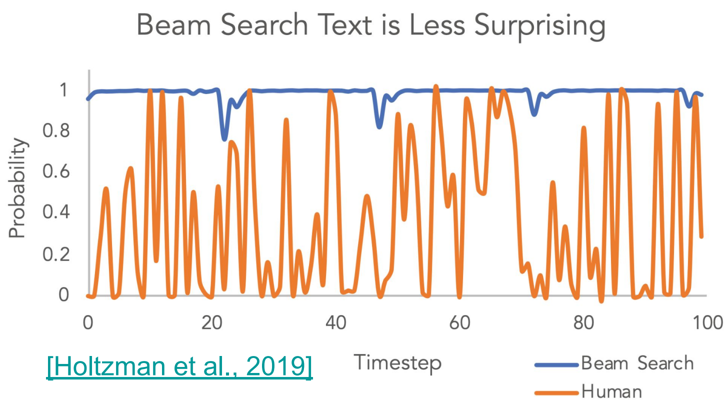
	Greedy	Beam Search	Top-k Sampling	Top-p (Nucleus) Sampling
At each step	Pick the best word	Try a few best words	Random sample from top-k	smallest set with cumulative probability > p
Output	One sequence	Several partial sequences	One sequence	One sequence

Weakness
Repetition as always select the most frequent token



Output Generation: Beam Search

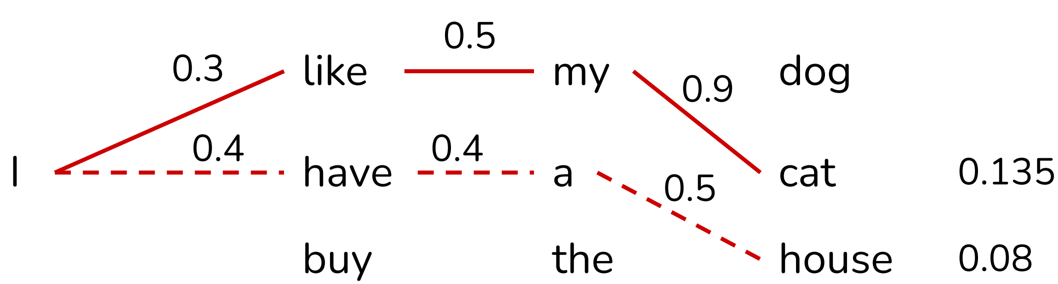
Greedy	Beam Search	Top-k Sampling	Top-p (Nucleus) Sampling
	Try a few best words	Random sample from top-k	smallest set with cumulative probability > p
	Several partial sequences	One sequence	One sequence



[Holtzman et al., 2019]

Common beam: 5 or 10

Weakness
Short sentences
Less diversity

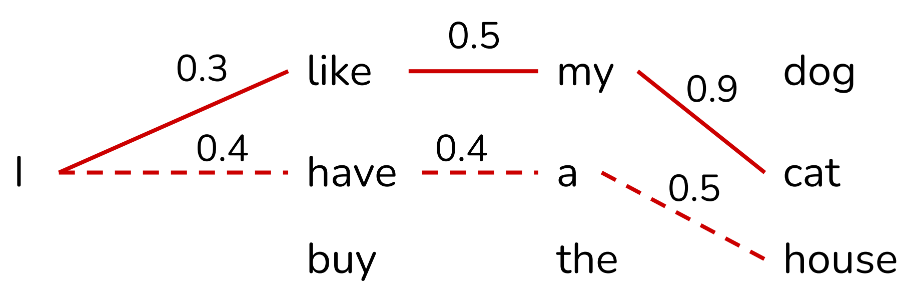


Output Generation: Top-k Sampling

	Greedy	Beam Search	Top-k Sampling	Top-p (Nucleus) Sampling
At each step	Pick the best word	Try a few best words	Random sample from top-k	smallest set with cumulative probability > p
Output	One sequence	Several partial sequences	One sequence	One sequence

Common k: 5, 10, 20

Note
k=1 → Greedy algorithm
↑k → more diverse/risky
↓k → more generic/safe



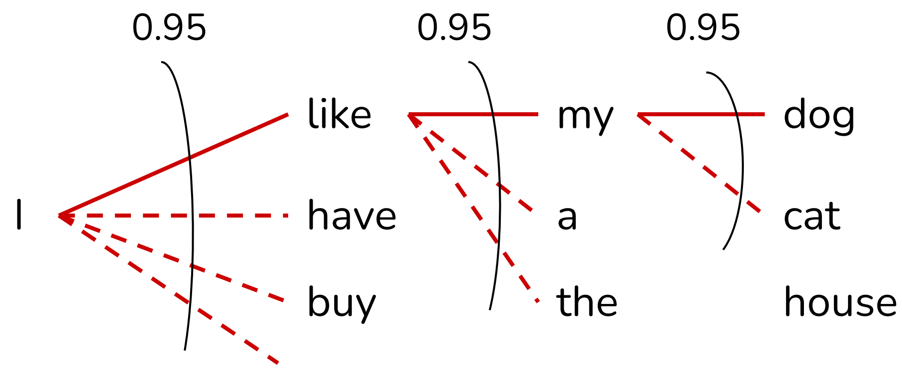
[Holtzman et al., 2019]

Output Generation: Top-p (Nucleus) Sampling

	Greedy	Beam Search	Top-k Sampling	**Top-p (Nucleus) Sampling**
At each step	Pick the best word	Try a few best words	Random sample from top-k	smallest set with cumulative probability > p
Output	One sequence	Several partial sequences	One sequence	One sequence

Common p: 0.95

Weakness
May not include surprising words



[Holtzman et al., 2019]

Output Generation: Decoding Strategies

TECHNISCHE UNIVERSITÄT DARMSTADT

	Greedy	Beam Search	Top-k Sampling	Top-p (Nucleus) Sampling
At each step	Pick the best word	Try a few best words	Random sample from top-k	smallest set with cumulative probability > p
Output	One sequence	Several partial sequences	One sequence	One sequence

Training

4-step recipe
[Fung et al., 2020]

1. Data
2. Model
3. Training
4. Evaluation

- Teacher forcing: Maximum Likelihood Estimation (MLE)
    - Maximize the conditional probability of target sequence
- Unlikelihood training [Welleck et al., 2020]
    - Minimize likelihood of undesired tokens

What makes a good conversation ?

Human judgment of conversational aspects

Avoiding Repetition	internal repetition; repetition across responses; partner repetition
Interestingness	interesting response: knowledge, engagingness
Making sense	coherent response
Fluency	grammatically correct
Listening	response related to user's utterance
Inquisitiveness	response and ask information about user

[See et al., 2019]

Some Methods to Make a Good Conversation

Input x → Encoder-Decoder Framework → Output y

- Large pretrained models: BART, T5, GPT-1/2/3
  → more **fluent** & **diverse response** thanks to large scale pretraining

- Decoding strategies
  - Top-k sampling, top-p (nucleus) sampling
    → reduce repetition
  - Guided decoding

- Input modification
  - Integrating attribute description

- Type embeddings
  - Using learned attribute embeddings

- Reinforcement learning

Large Scale Pretraining

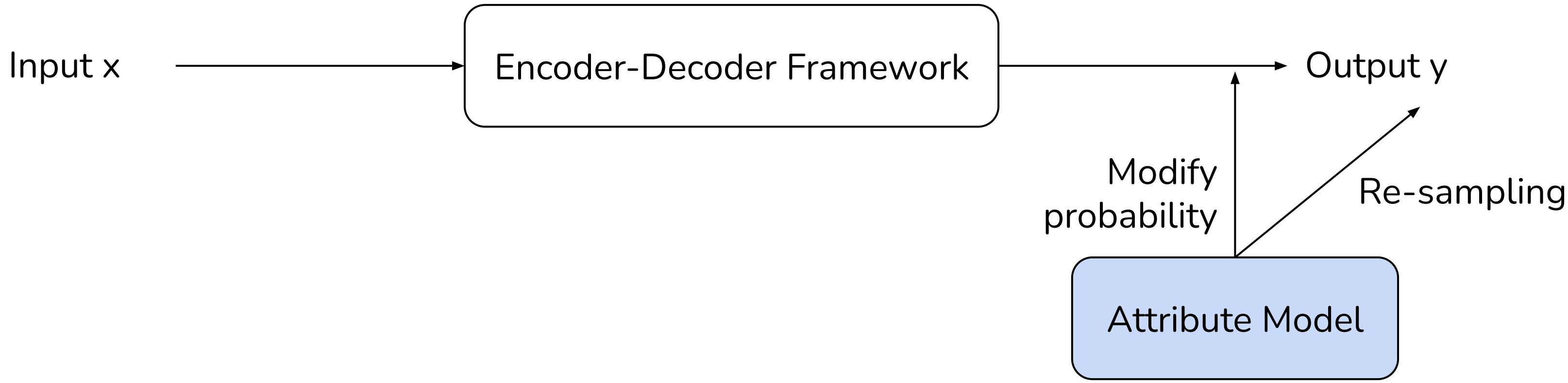
TECHNISCHE UNIVERSITÄT DARMSTADT

Input x → Encoder-Decoder Framework → Output y

Large pretrained models: BART, T5, GPT-1/2/3 → proposed for general text generation

PLATO	DialoGPT	Meena	BlenderBot	TOD-BERT
Twitter, Reddit (En)	Reddit (En)	public domain social media conversations (En)	Reddit discussions (En)	9 task-oriented datasets (En)
BERT	GPT-2	Evolved Transformer	Poly-encoder Transformer + Seq2seq	BERT
- Response generation - Latent act recognition	- Response generation - Maximum Mutual Information	- Minimize perplexity of next token	- Masked LM - Ranking for retrieval - Response generation - Retrieve & refine - Unlikelihood training	- Masked LM - Response contrastive loss

Guided Decoding



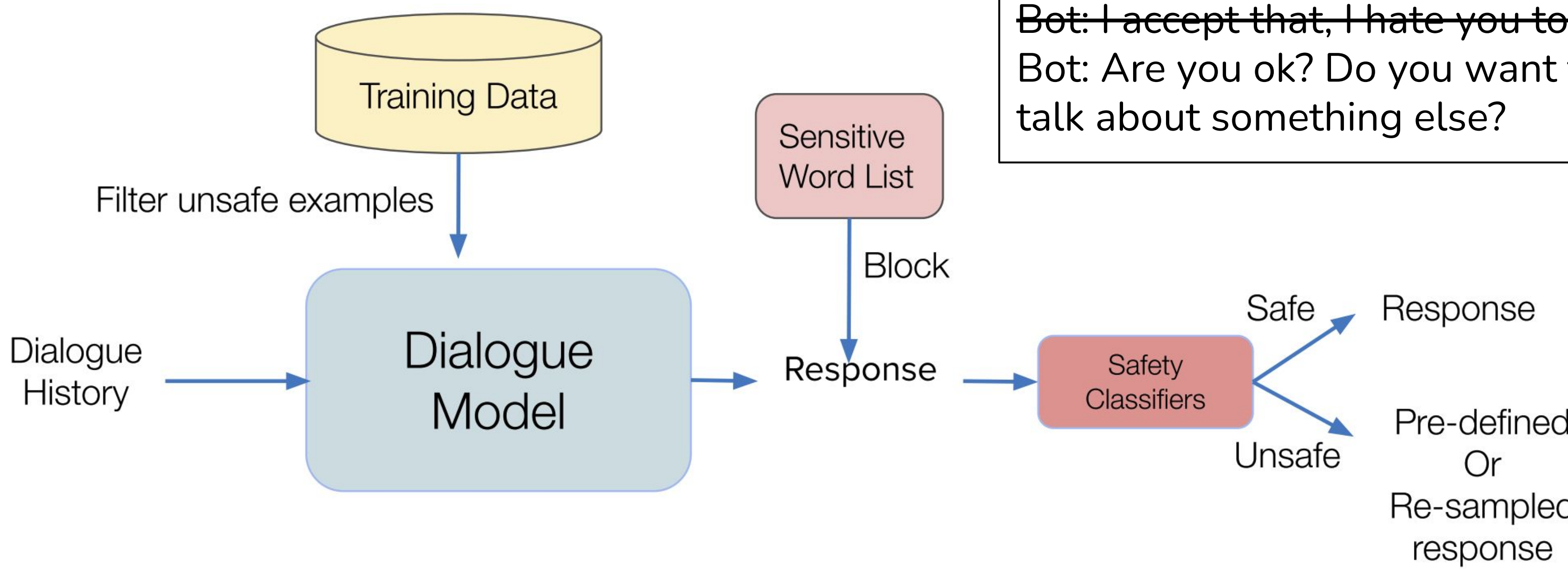
1. Define an attribute model to score the generated sequence
2. Guide the decoding process
    - Re-sampling if not satisfy attribute guide
    - Modify the probability distribution with attribute scores [Madotto et al., 2020]

Guided Decoding

[Safety in Open-domain chatbots](#)

Dialogue systems trained on large scale data may inherit biases from such data
→ may produce toxic, harmful and biased responses
→ bring bad experience to user

Human: I hate everyone. Acceptable?
~~Bot: I accept that, I hate you too.~~
Bot: Are you ok? Do you want to talk about something else?



[Fung et al., 2020]

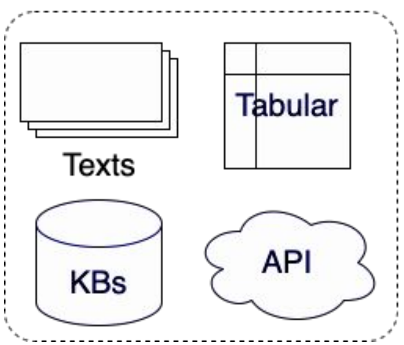
Input Modification

Input x → Encoder-Decoder Framework → Output y

+
Attribute description

E.g., Dialog history + [positive]; [sad] + Dialog history

Knowledge

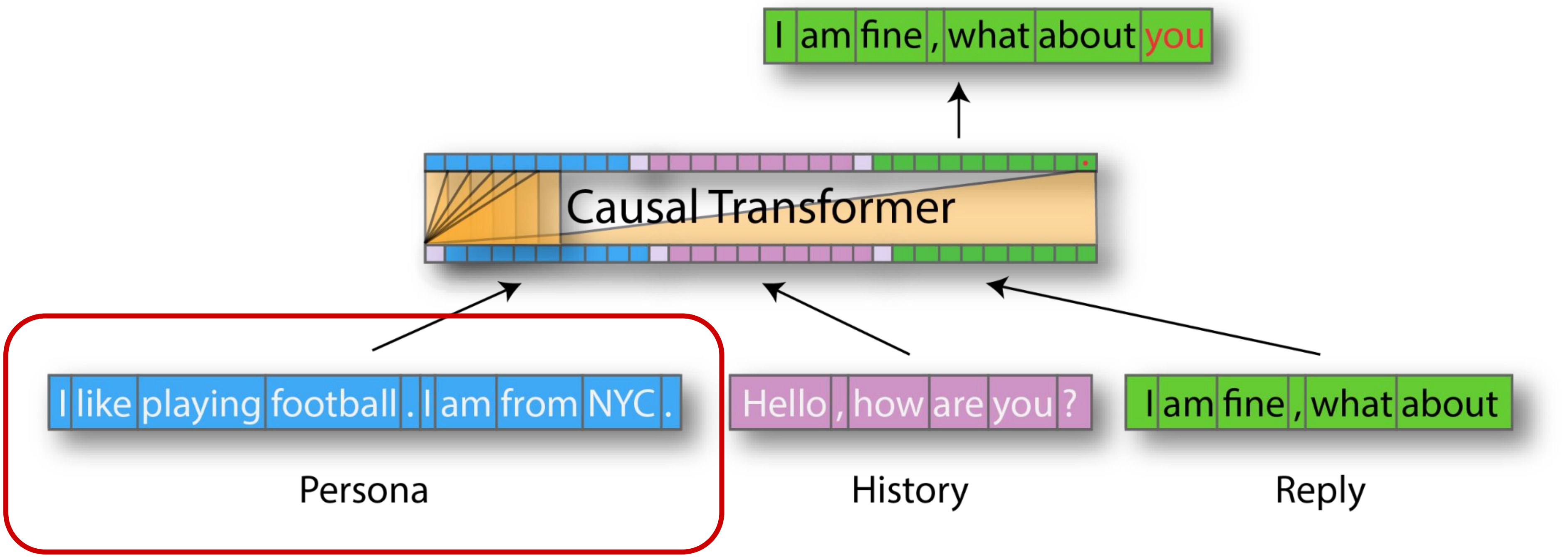


- Retrieval methods:
    - IR systems (TF-IDF, BM25)
    - Neural retriever: dense vectors
    - Generating API query
- Knowledge to text → add to input

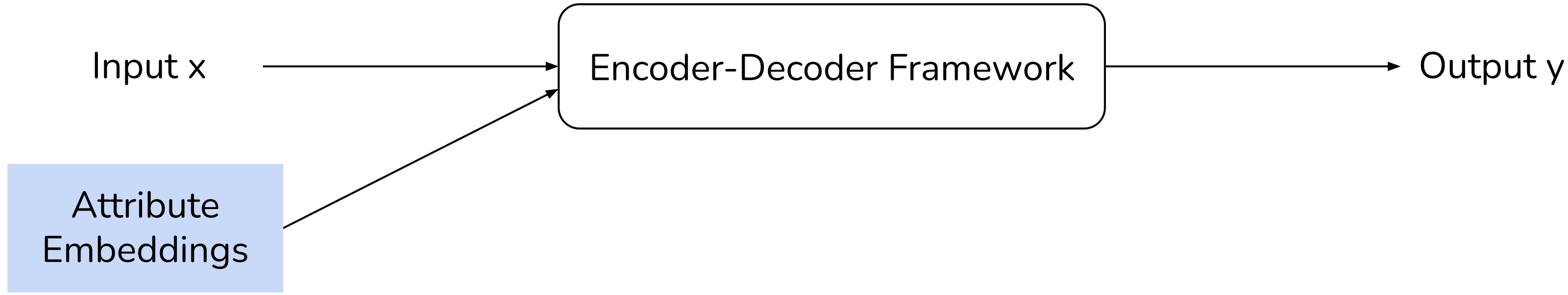
  E.g., “Input” + [restaurant] Sushi Bar

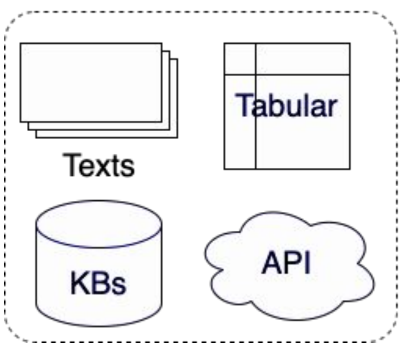
Input Modification

Personalization: TransferTransfo Model



Attribute Embeddings

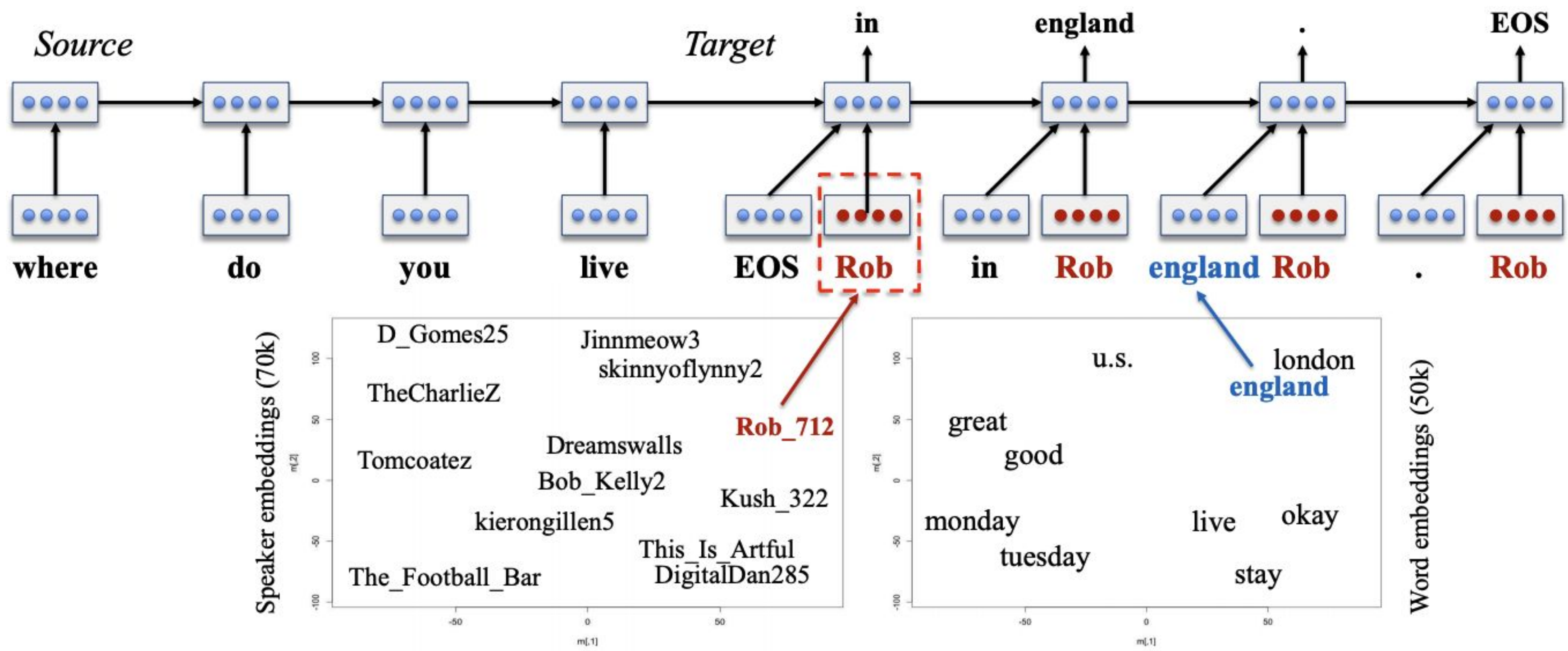




- Embeddings of the target attributes
    - Learned during training the conversational model
    - Obtained from additional methods such as clustering, link prediction for knowledge base integration

Attribute Embeddings

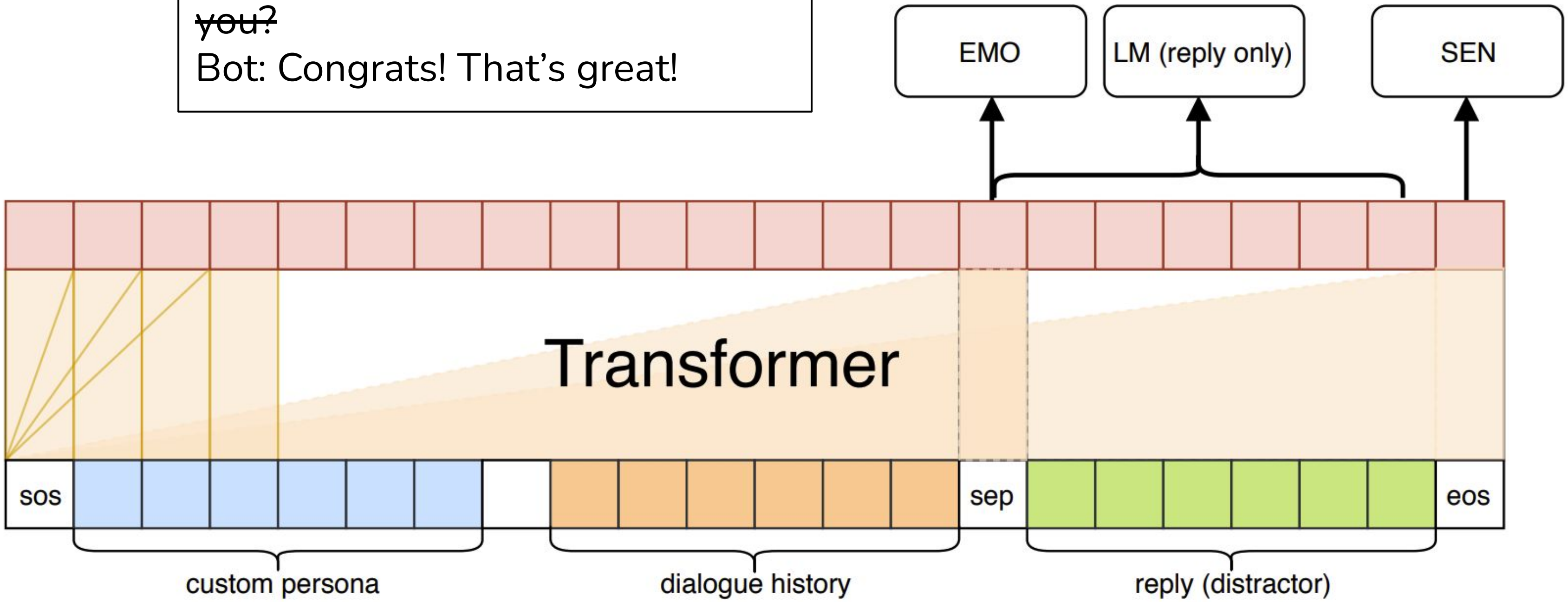
Personalization: Speaker Model



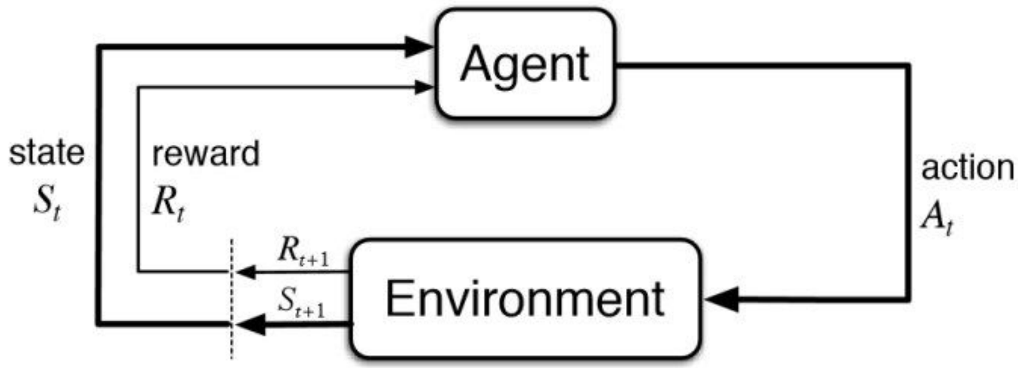
Attribute Embeddings

Empathy: CAiRE

Human: I finally got promoted.
~~Bot: Why would anyone promote you?~~
Bot: Congrats! That's great!



Reinforcement learning



- Reinforcement learning
  - Cast a text generation model as markov decision process
    - *State*: dialog history + previous generated tokens
    - *Actions*: possible tokens
    - *Policy*: conversational model + decoding strategy
    - *Rewards*: attribute models for a good conversation
      - Politeness, sentiment, ..

Reinforcement learning

Personalization:

- Reward function: capture consistency between a response and persona facts
  - Persona consistency
  - Topical coherence
  - Fluency
  - Repeated tokens

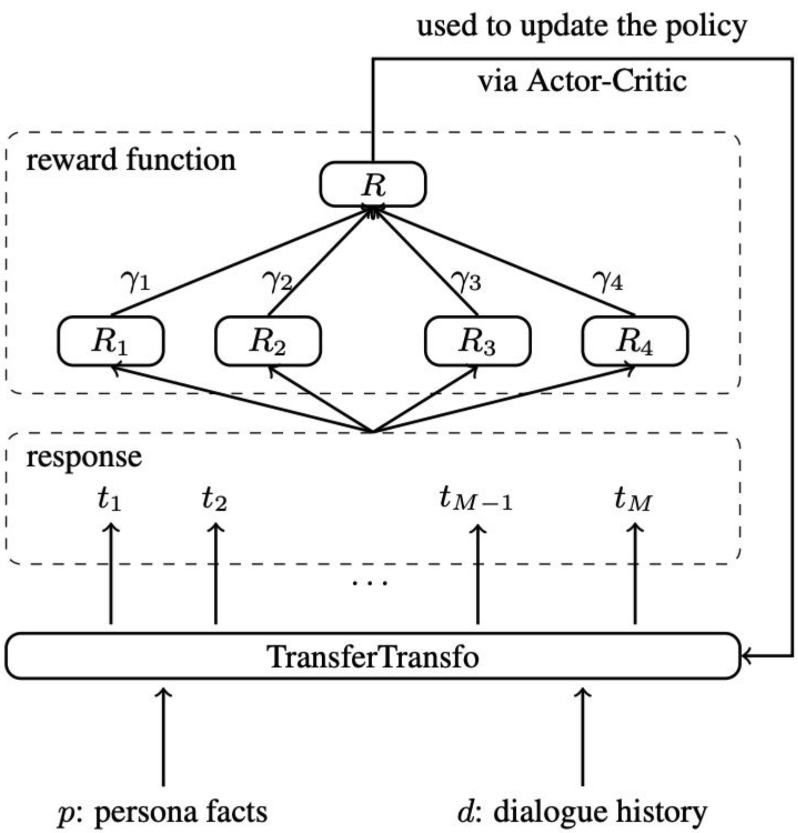


Figure 1: An abstract view of our RL approach.

[Mesgar et al., 2021]

Evaluation

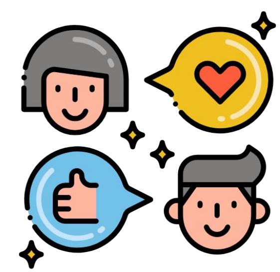
4-step recipe
[Fung et al., 2020]

1. Data
2. Model
3. Training
4. Evaluation

System: The brown fox jumps

Reference: The fox

Automatic Evaluation



Human Evaluation

[Ji et al., 2020]

Automatic Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Compare with reference response

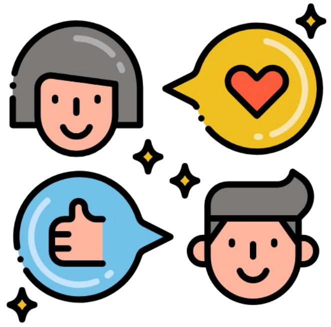
System: The brown fox jumps

Reference: The fox

- Main categories
  - Perplexity: how likely a model generate the reference response
  - N-gram based overlap: BLEU, ROUGE-L
  - Distinct N-gram: diversity
    → Weakness: surface level, correlate poorly with human judgement
  - Model based metrics: BERTScore [Zhang et al., 2020], Adversarial Success [Kannan & Vinyals, 2017; Li et al., 2017]
    → Weakness: not interpretable, not always align with human judgement

Human Evaluation

- Interaction setup
  - Dialogue history, gold response, generated response
  - Directly interact with systems
- Evaluation setup
  - Likert: give ratings according to some criteria, e.g., fluency, consistency, factual etc.
  - Selection preference: select one system among presented systems (usually btw 2)
- Weaknesses
  - Expensive & time consuming
  - Difficult quality control, inconsistency in evaluation



Summary: Evaluation

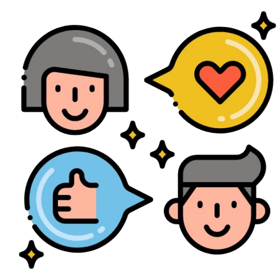
4-step recipe
[Fung et al., 2020]

1. Data
2. Model
3. Training
4. Evaluation

System: The brown fox jumps

Reference: The fox

Automatic Evaluation



Human Evaluation

- Improper or offensive language
- Factual consistency

[Ji et al., 2020]

Summary

- A lot efforts have been made
- But still many **improvements** ahead in Conversational AI
- Evaluation remains a huge challenge
  - Need better ways of automatic evaluation
- **Most exciting areas** of NLP!