




# Proactive Conversational Agents

Lizi Liao (Singapore Management University)
Grace Hui Yang (Georgetown University)
Chirag Shah (University of Washington)

# Who we are

Lizi Liao

Singapore Management University

Grace Hui Yang

Georgetown University

Chirag Shah

University of Washington

WARNING: this is an emerging research area, conclusions in this tutorial may be out-of-date soon!

# Outline

Part-1

- **Introduction**
- Interactive exercise-1
- Overview of proactive conversation agent

Part-2

- Proactive ontology expansion
- Learning to ask & topic shifting
- Counterfactual utterance generation

Part-3

- Response quality control
- Interactive exercise-2

# Brief History of Conversational Agents

Rule-based Era

Machine Learning Era

Deep Learning Era

MIT Eliza
*Pattern+Transform*

Rule+ML
*Mixed Systems*

Rule+ML
*Smart Assistant*

Transformer
*Big model+Big data*

Switch Transformer
Wu Dao 2.0
*1.75 trillion parameters*

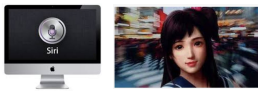
LaMDA
*137B parameters*

2022-today

1966－1994

- *MIT Eliza*
- *ALICE rule-based chatbot*

2011－2014



- *Apple Siri*
- *Microsoft Xiaobing*

2016－2018


- *Google's personal assistant*
- *Amazon Alexa*
- *-Tianmao's Jingling*

2020


OpenAI

*-Large-scale pretrained models*
*Google: Meena*
*OpenAI: GPT-3*

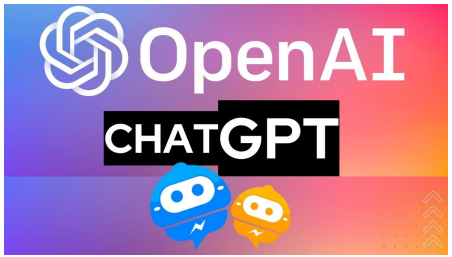
2021

- *Tesla bot*
- *Xiaomi's Cyberone*
*Human-like generative conversational model*

- *Google LaMDA*
- *ChatGPT*
- *Google Bard*
- *Earnie Bot*

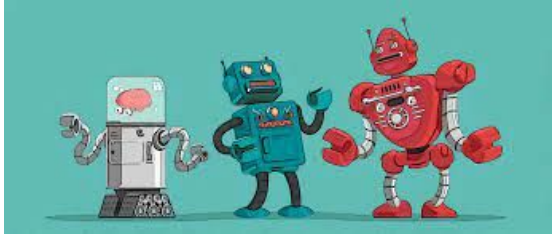
[Huang et al., 2022]

# Hot Types in 2022/2023

| | | |
|---|---|---|
| Solving Tasks |  <br>ChatGPT<br>(From OpenAI, 2022.12) |  <br>Claude<br>(From Anthropic AI, 2023.1) |
| Personification |  <br>Character.AI<br>(From Character, 2022.9) | LING XIN INTELLIGENCE<br>聆心智能<br>乌托邦<br>AI topia<br>(From Ling Xin Intelligence, 2022.11) |

[Huang et al., 2022]

# Why we need? – Solving Tasks



“Can you help me book a restaurant for two?”

“Please book a flight to Singapore on this Monday morning.”

“When and where is the WSDM 2023 conference?”

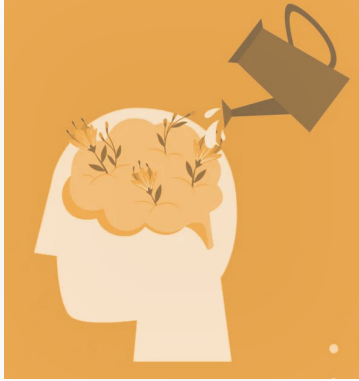
“I would like to know more about ChatGPT.”

“What letter comes next in the following sequence: M T W T F __”

“Plan a two days tour to singapore.”

…

# Why we need? – Emotion Support



“It is raining outside. I am kind of feeling lonely.”

“I hope to have chance talking to my grandma who passed away.”

“I would like to talk to Shakespeare.”

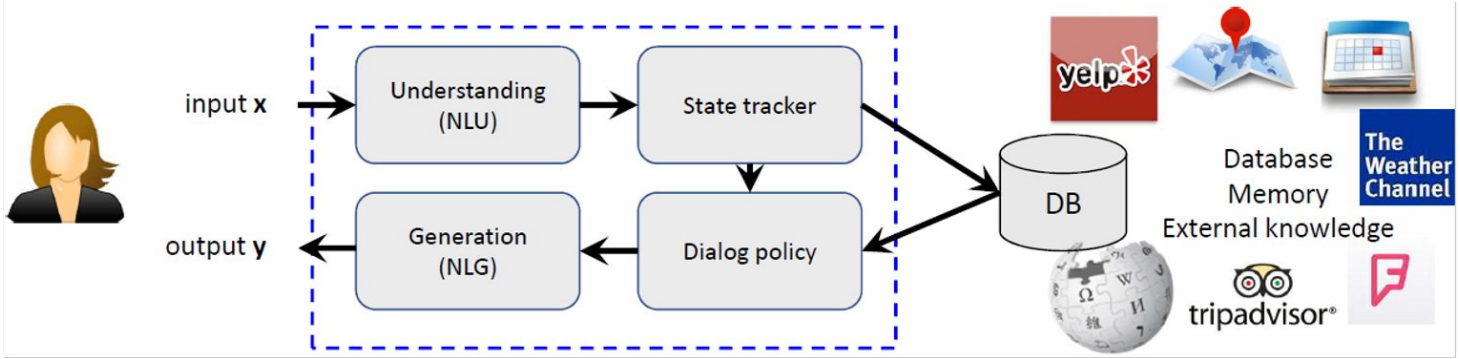
“I am preparing for my first show. Can you help me to built up my confidence?”

“I always feel upset and everything is meaningless. Why am I still here?”

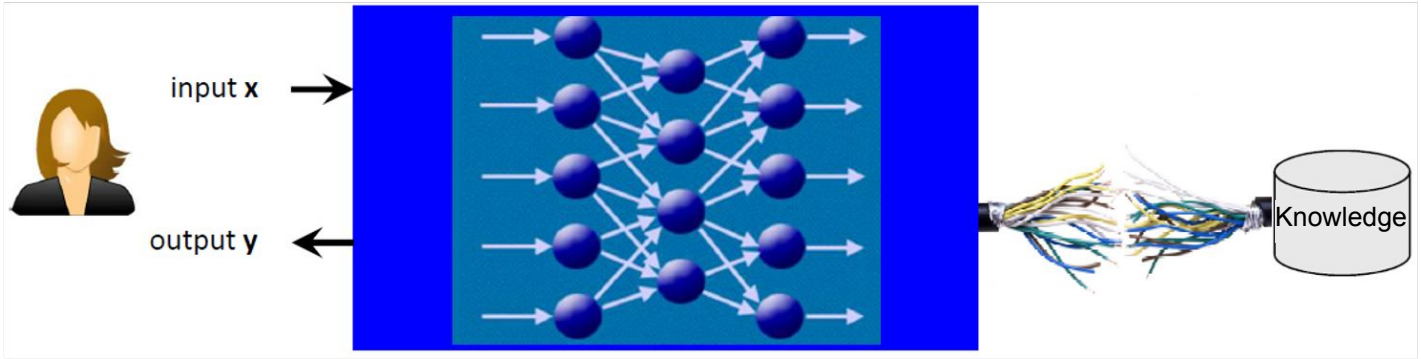
…

# Task-oriented Bot v.s. Open Chatbot

Task-oriented

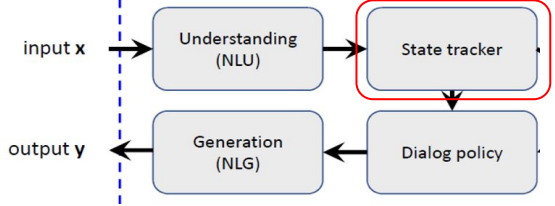



Open Chat




(Young et al., 2013; Tur & De Mori, 2011; Ritter et al, 2011; Sordoni et al. 2015; Vinyals & Le 2015; Shang et al. 2015)

# Typical State Tracking Module




*Find a good eating place for chinese food*

O O B-rating O O O B-type O

*Find a good eating place for chinese food*

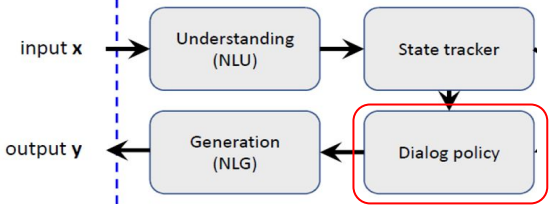

| SLOT | VALUE |
|---|---|
| Rating | *good* |
| Type | *chinese* |
| # of people | ? |
| Area | ? |

Intent Classification

Sequence Labeling

FIND_RESTAURANT(rating="good", type="chinese")

# Typical Dialogue Management Module

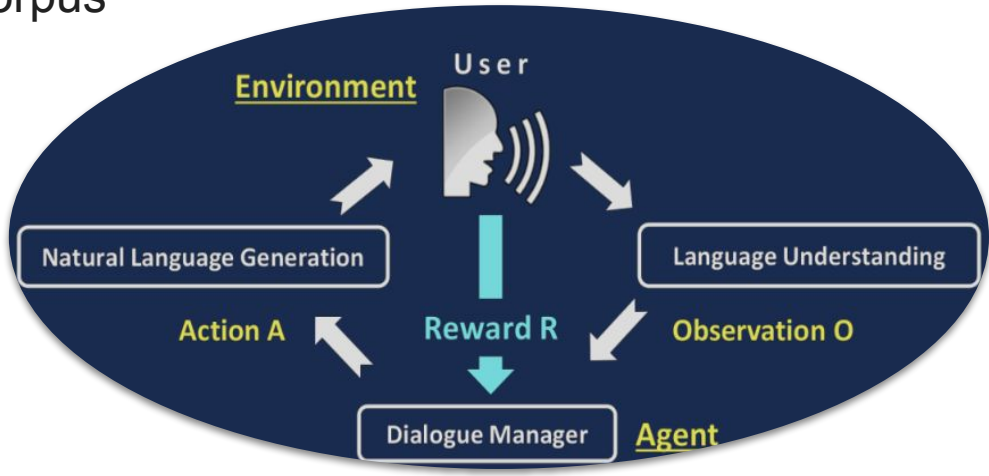



- **Rule-based**
  - Huge hand-crafting effort
  - Non-adaptable and non-scalable
  - But it works

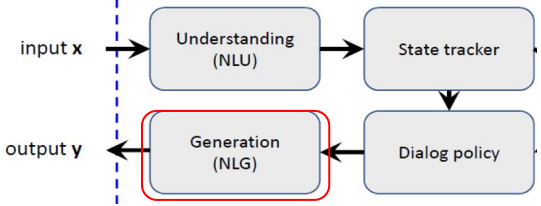
FIND_RESTAURANT(rating="good", type="chinese")

REQUEST(area, price_range)

- **Supervised**
  - Learn to 'mimic' answers from a corpus
  - Assumes optimal human behavior
  - Short or long-term planning
- **Reinforcement learning**
  - Try-and-error
  - Long-term planning



# Typical Generation Module




- Meaning representation ⟹ natural language utterances

***Dialogue act***

REQUEST(area, price_range)

***Realisations***

Where do you want to go? Any price range requirement?

Do you have any specific location in mind? What price range?

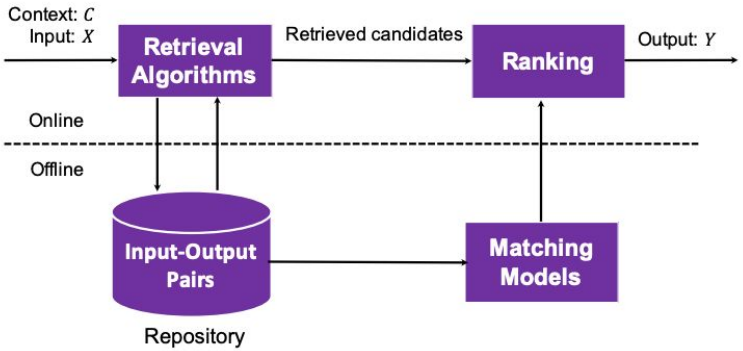
1. Template-based
2. Statistical approaches

   [Walker et al.,2002; Stent et al., 2009; Sethlefs et al., 2013; Cuayahuitl et al., 2014]
3. Seq2Seq models

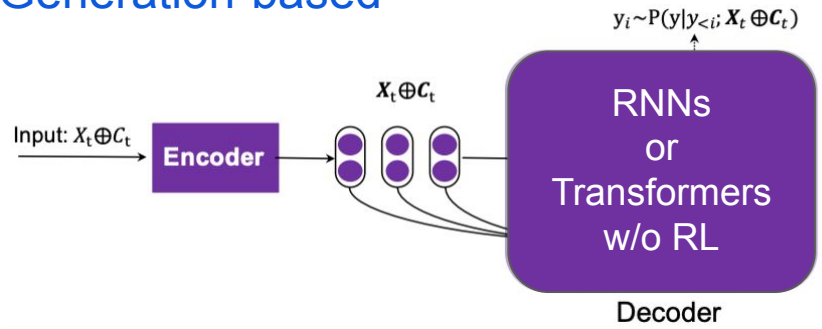
   [Wen et al.,2015; Mei et al., 2016; Tran and Nguyen, 2017; Guu et al., 2017, Fedus et al., 2018]

# Open-chat Generation Models

## ❑ Retrieval-based



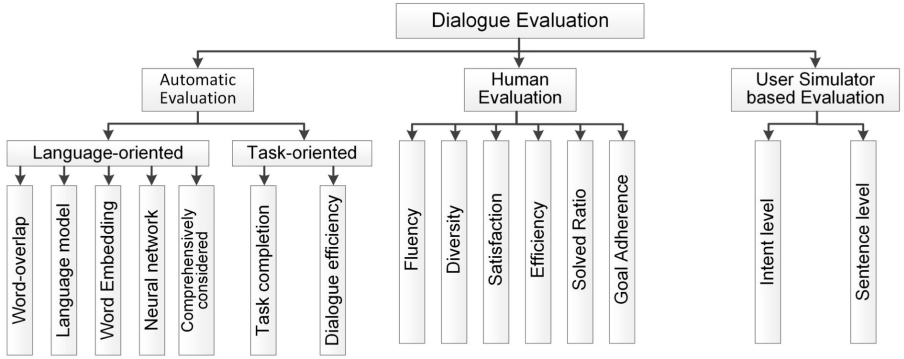

## ❑ Generation-based




[Huang et al., 2020]

| | |
|---|---|
| Semantics Issue | |
| User | I come from Shanghai. |
| Machine | Cool! I am from Beijing, we come from the same city! |
| User | Back to Beijing now. It is so cold but I am very happy! |
| Machine | When will you come to Beijing? |
| Consistency Issue | |
| User | Are a boy or a girl? |
| Machine | I am a boy. |
| User | Are you a girl? |
| Machine | Yes, I am a girl. |
| Interactiveness Issue | |
| User | Terrible breakfast today. |
| Machine | How about together tomorrow morning? |
| User | Sure. Would be great with your recommendation. |
| Machine | Really? |

# How to Evaluate?




[Xinmeng Li et al., 2021]

## ❑ Subjective: Human judgement

- ➢ Adequacy: correct meaning
- ➢ Fluency: linguistic fluency/naturalness
- ➢ Coherence: fluency in the dialogue context
- ➢ Variation: multiple realisations of the same meaning
- ➢ …

## ❑ Objective: Automatic evaluation metrics

- ➢ Word overlap: BLEU, METEOR, ROUGE
- ➢ Embedding-based: greedy matching, embedding average
- ➢ Task-oriented: item error rate, success rate, entity F1
- ➢ …

Big gap and negative correlation between human judgements and automatic measure.
Real user trial is still the best way to evaluate.

# Emerging Trends - ChatGPT

- On Nov 30, 2022, Open AI released ChatGPT
  - GPT stands for “Generative Pre-Trained Transformer”
    - Generative: a model that can generate content
    - Pre-trained: it leverages large scale ML model trained over large scale data and the model can be stored and continue to be trained
    - Transformer: a type of neural network architecture
  - It is built on the 3rd generation of GPT model (GPT-3.5) released by Open AI
    - GPT-2 is free and anyone can download and play with it
- Over 100 million users in January 2023
- Many rumors, concerns, and reality
  - Ending of Google Search (and probably Amazon Alexa etc)
  - Concerns in academic and education
  - Big layoffs in Silicon Valley
  - Revolutionary language/content generation (do we still need writers, journalist, musicians, painters, movie directors?)
  - Revolutionary data analysis and synthesis (do we still need data analysts, secretary, lawyers?}
- On Feb 6, 2023, Google released Bard
- On Feb 7, 2023, Microsoft announced New Bing with ChatGPT integration

# Outline

Part-1

- Introduction
- **Interactive exercise-1**
- Overview of proactive conversation agent

Part-2

- Proactive ontology expansion
- Learning to ask & topic shifting
- Counterfactual utterance generation

Part-3

- Response quality control
- Interactive exercise-2

# Exercise instructions

- Create pairs.
- One person acts as the user and the other acts as the agent.
- The user is given a task. **Don’t reveal it to the agent!**
- The user starts asking questions to the agent to fulfil their task.
- The agent can use **web search** or other **online means** to find the information, but **can’t provide any information or suggestions that don’t exist online**.
- Each response should take no more than **30 seconds**.
- Spend no more than 5 minutes on this and then reverse the roles. Repeat.
- Debriefing: 5 minutes.

# Outline

Part-1

- Introduction
- Interactive exercise-1
- **Overview of proactive conversation agent**

Part-2

- Proactive ontology expansion
- Learning to ask & topic shifting
- Counterfactual utterance generation

Part-3

- Response quality control
- Interactive exercise-2

# On the Proactiveness – Handling the unseen

- New Intent

  Can you show your vaccination QR code to me?

  Which vaccination have you taken? How many doses?

  *Request_VacCert*

  *Request_VacType*
  *Request_dose*

- New Value

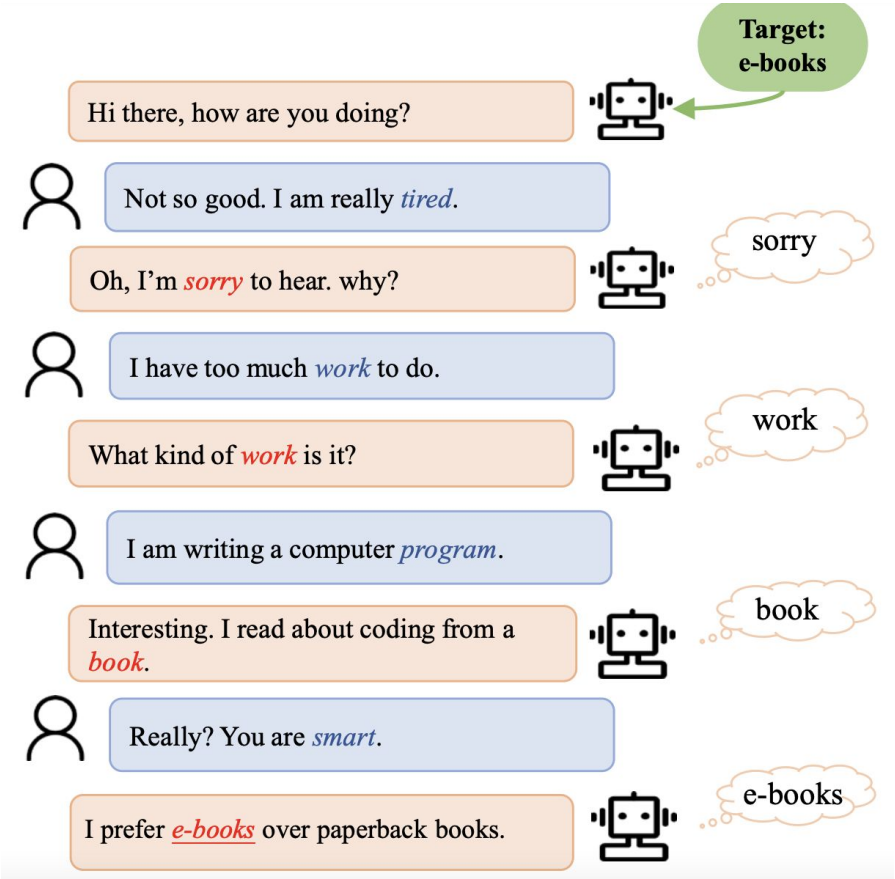
  How is Vaxzevria? What type is it?

  Can you book a table for me in TAI ER @ Suntec?

- New Slots

| Vac_Type | Dosage | PCR_result |
|---|---|---|
| inactivated vaccines | one | positive |
| mRNA vaccines | two | negative |
| live-attenuated vaccines | three | unknown |
| … | … | … |

# On the Proactiveness – Driving the conversation




- Asking Clarifying Questions
- Suggest Useful Questions in search
- Mix chit-chat with task-oriented interactions
- Mix-initiative in control flow
- Topic shifting

…

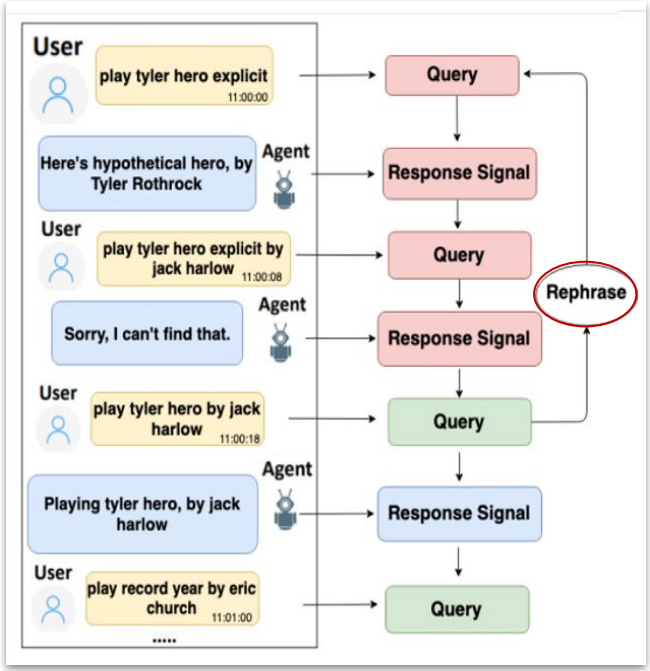
Jianheng Tang, et al. "Target-Guided Open-Domain Conversation." *ACL* 2019.

# On the Proactiveness – Response control with human feedback

- Enable learning autonomously from user-system interactions (e.g. barge-in, reformulations), system signals, and predictive models
  - Explicit feedback
    *e.g. "did I answer your question?" "yes"*
  - Implicit feedback
    *e.g. user barge-in a turn or rephrase her request*
  - Unsolicited feedback
    *e.g. "thank you" or "I am not Derek, I am Dave"*

- Defect correction with self-learning, rephrase detection, failure point isolation
  - Precomputed rewriting
  - Online rewriting (long tail situations)

- Reinforcement learning for long-term



# Outline

Part-1

- Introduction
- Interactive exercise-1
- Overview of proactive conversation agent

Part-2

- **Proactive ontology expansion**
- Learning to ask & topic shifting
- Counterfactual utterance generation

Part-3

- Response quality control
- Interactive exercise-2

# Proactive Ontology Expansion

## ➢ New Intent Discovery

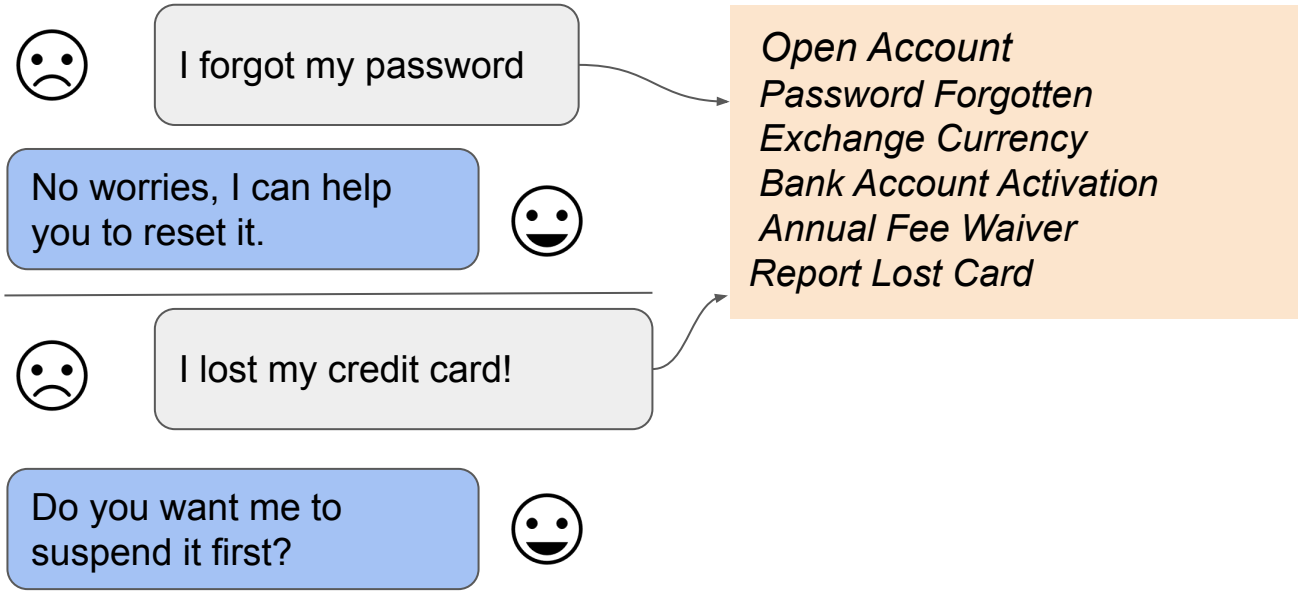
- Definition
- Discover from scratch (all intents are unknown)
- Discover new from data (some intents are known)

## ➢ New Slot Induction

- Definition & Traditional Approaches
- Emerging Methods

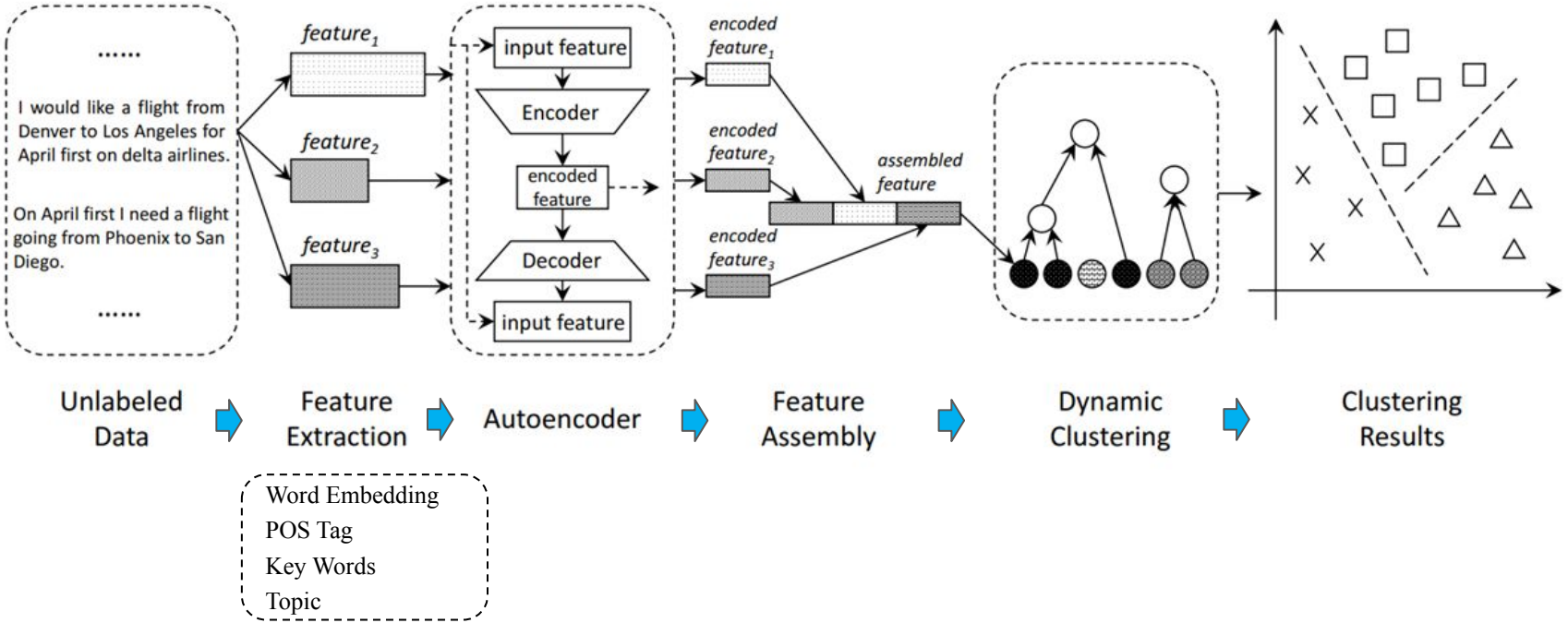
# Intent Discovery

Intent Discovery aims to uncover novel intent categories from user utterances. It is a critical task for the development of a practical dialogue system.



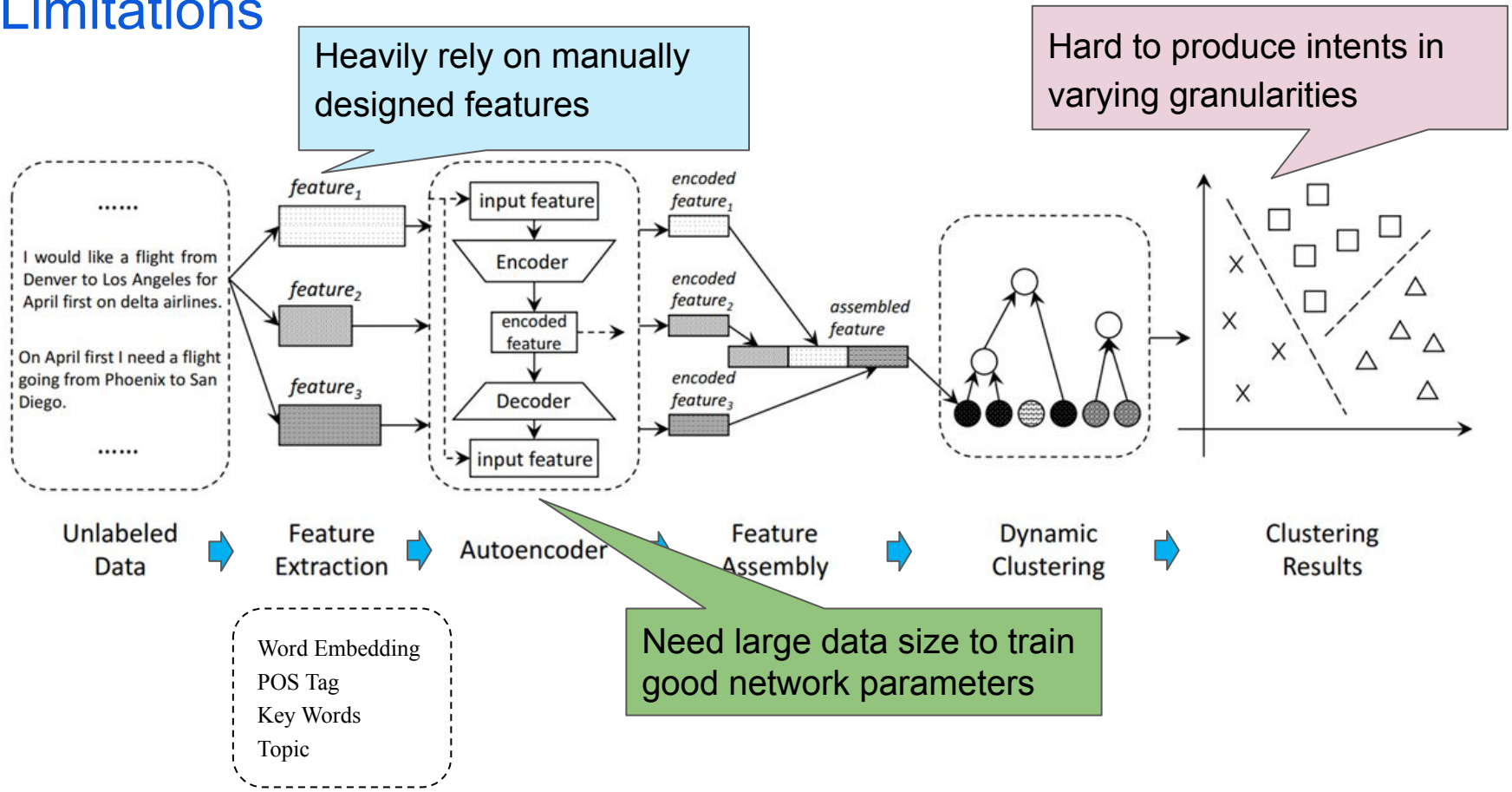
# Discover from scratch

- Classes are unknown, use clustering to group utterances of similar intents
- Cluster assignments used as new intent labels or as heuristics for annotations




Shi, Chen et al. “Auto-Dialabel: Labeling Dialogue Data with Unsupervised Learning.” *EMNLP*(2018).

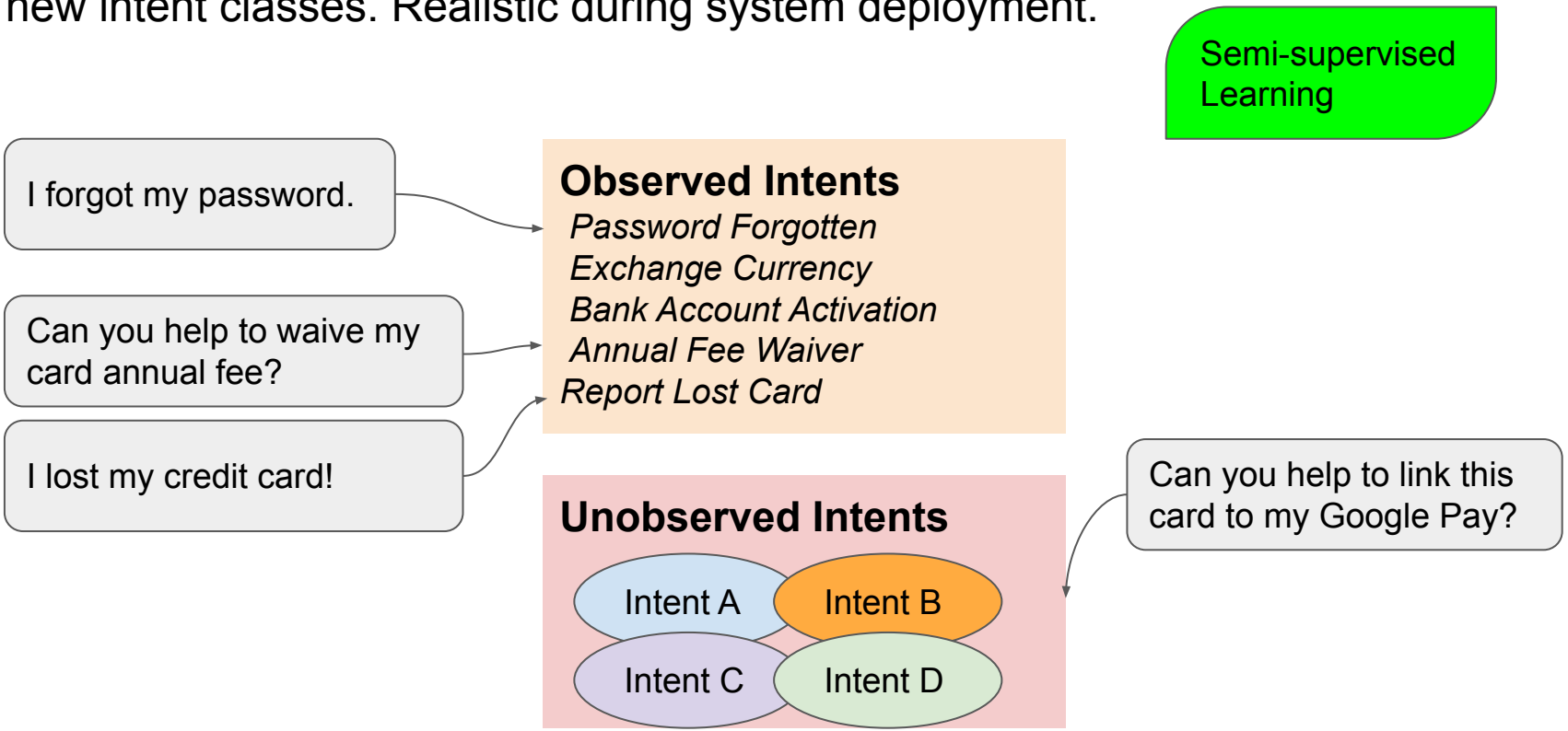
# Limitations




Shi, Chen et al. “Auto-Dialabel: Labeling Dialogue Data with Unsupervised Learning.” *EMNLP*(2018).

# Discover new with known intents

- Part of the dialogue data is labled with known intents. Other part contains new intent classes. Realistic during system deployment.



# Discover new with known intents

**Essential questions:**

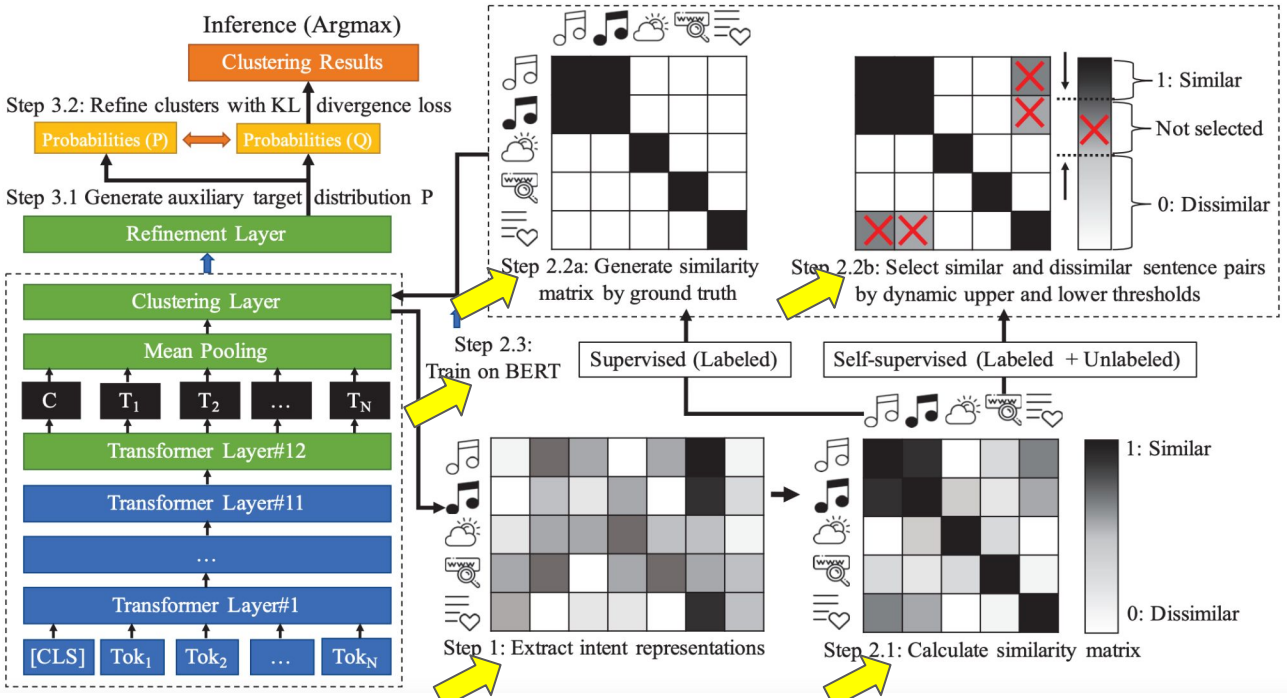
Semi-supervised Learning

***Q1. How to better cluster the utterances?***

***Q2. How to learn semantic utterance representations to provide proper cues for clustering?***

# For Q1. + Constrained Adaptive Clustering

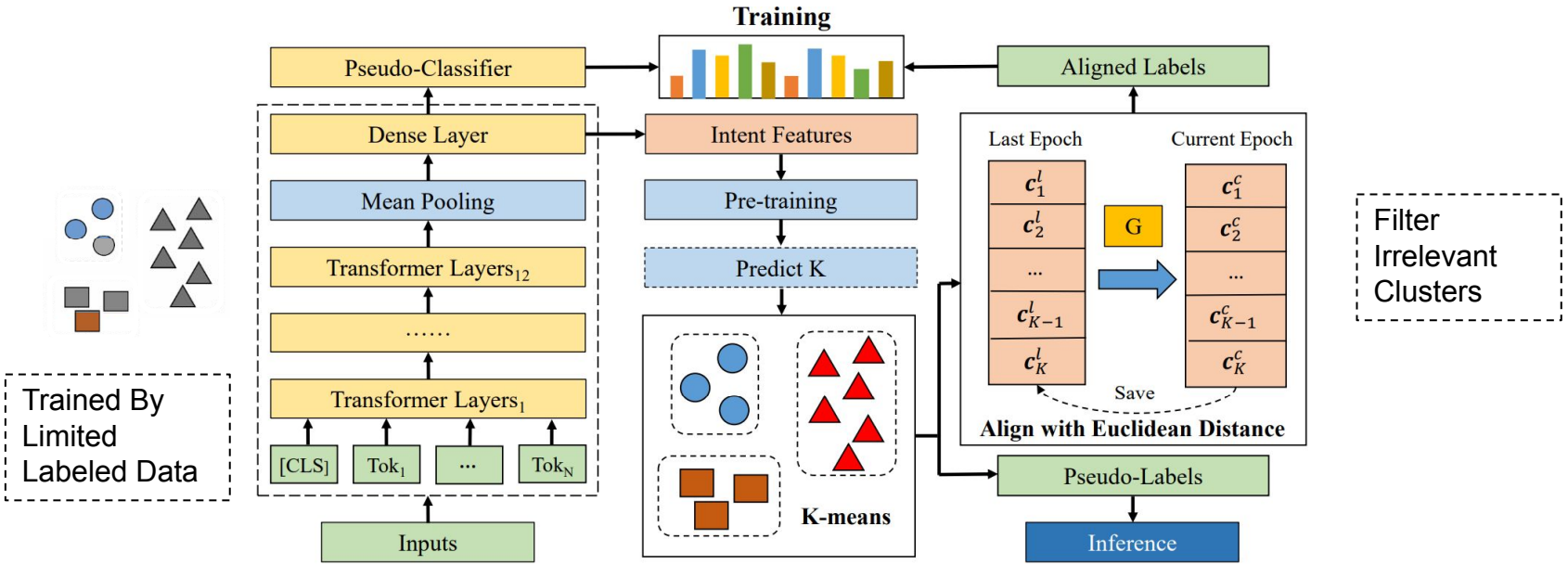
- Existing methods incorporate prior knowledge by intensive feature engineering
- Incorporate pairwise constraints as prior knowledge to guide the clustering




Ting-En Lin et al. "Discovering New Intents via Constrained Deep Adaptive Clustering with Cluster Refinement." *AAAI*(2020)

# For Q1. + Aligned Clustering

- Relying on Transformer to encode deep features
- Obtain pseudo supervised signals using K-means
- Align clusters between training epochs (preserve learned signals)




Zhang, Hanlei, et al. "Discovering new intents with deep aligned clustering." *AAAI*(2021)

# Discover new with known intents

**Essential questions:**

Semi-supervised Learning

***Q1. How to better cluster the utterances?***

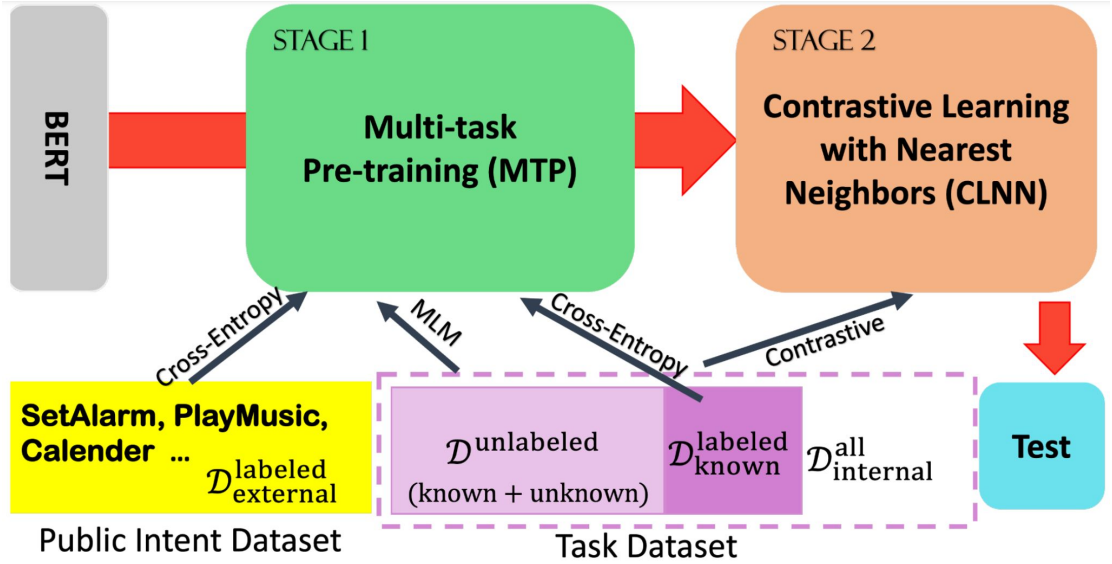
***Q2. How to learn semantic utterance representations to provide proper cues for clustering?***

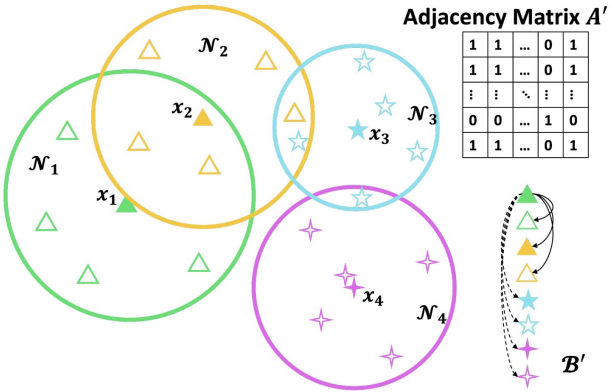
# For Q2. + Pre-training and Contrastive Learning

Directly apply vanilla pre-trained language model? ✗

Use labeled utterances of known intents to train encoder? ✗

- Multi-task pre-training using both external data and internal data
- Apply contrastive learning with nearest neighbors

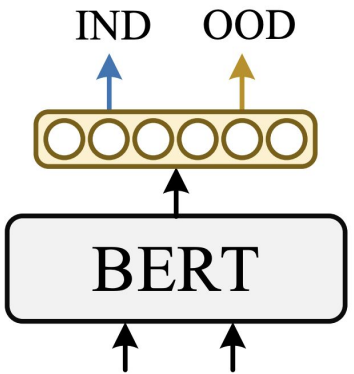


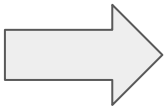



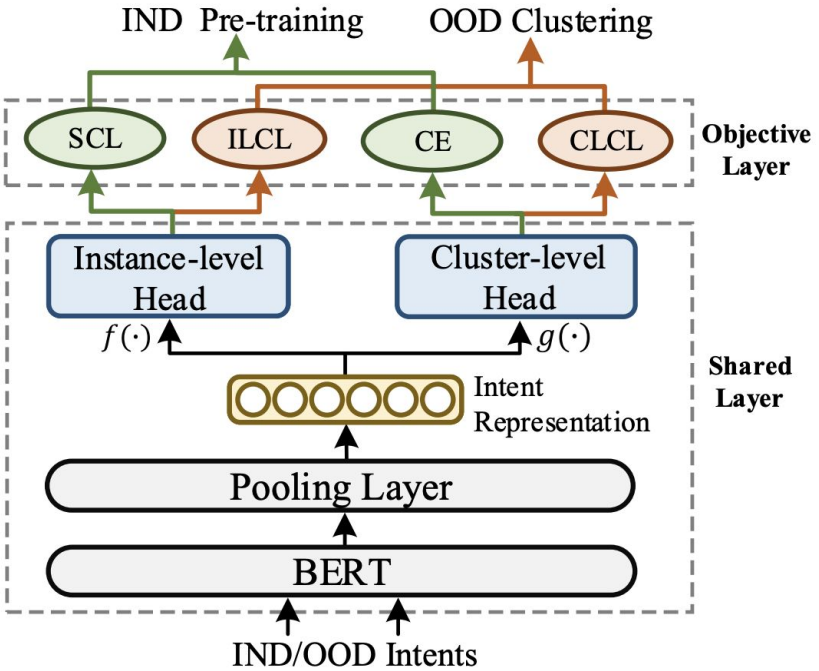

Zhang Yuwei, et al. "New Intent Discovery with Pre-training and Contrastive Learning." *ACL*(2021).

# For Q2. + Aligned Knowledge Transfer

- ➔ Pre-train an in-domain intent classifier, extract OOD intent representations by it, then perform clustering on extracted OOD intent representations
- ➔ Equip the traditional IND pre-training stage with a similar contrastive objective as the clustering stage



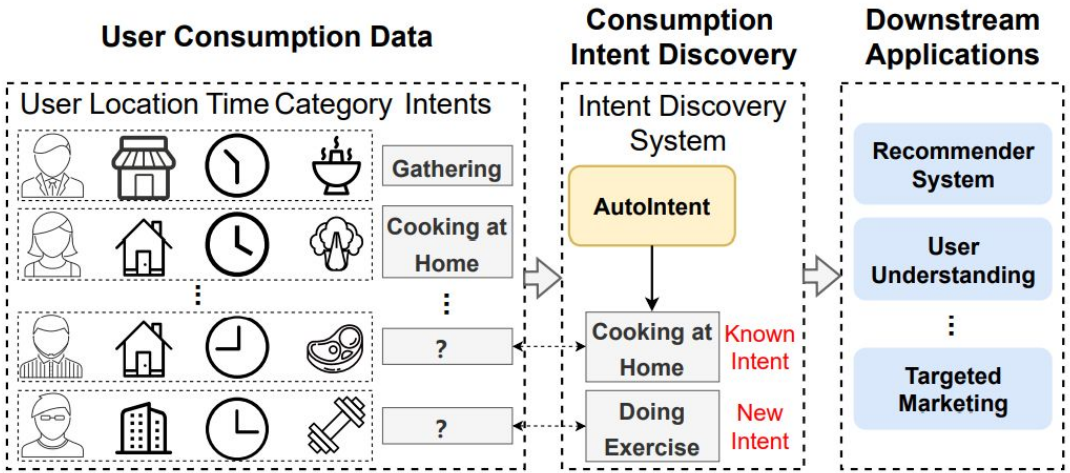






Mou, Yutao, et al. "Disentangled knowledge transfer for OOD intent discovery with unified contrastive learning." *ACL*(2022).

# For Q2. + Graph Neural Networks

- Discovering new consumption intents is crucial for downstream applications
- How to encode the consumption intent related to multiple aspects of preferences?
  – *Consumption is not only determined by users' intrinsic preferences like price and brand but also largely affected by spatial and temporal factors*

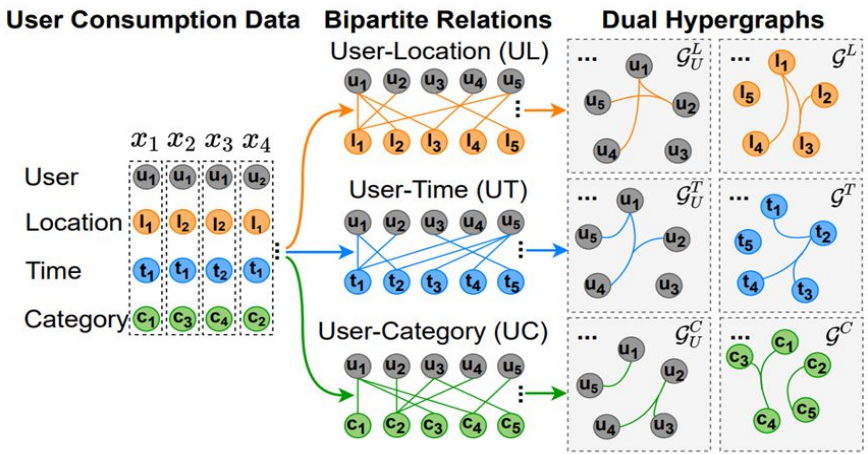


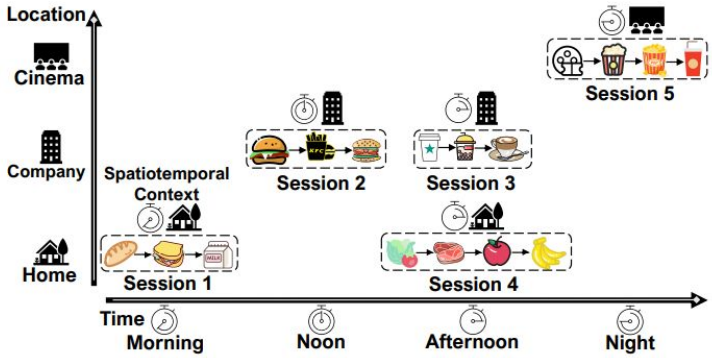


Li, Yinfeng, et al. "Automatically Discovering User Consumption Intents in Meituan." *KDD*(2020).
Li, Yinfeng, et al. "Spatiotemporal-aware Session-based Recommendation with Graph Neural Networks."*CIKM*(2022).

# For Q2. + Graph Neural Networks

- construct three groups of dual hypergraphs to capture relations under aspects
- utilize the hypergraph neural networks to extract disentangled intent features



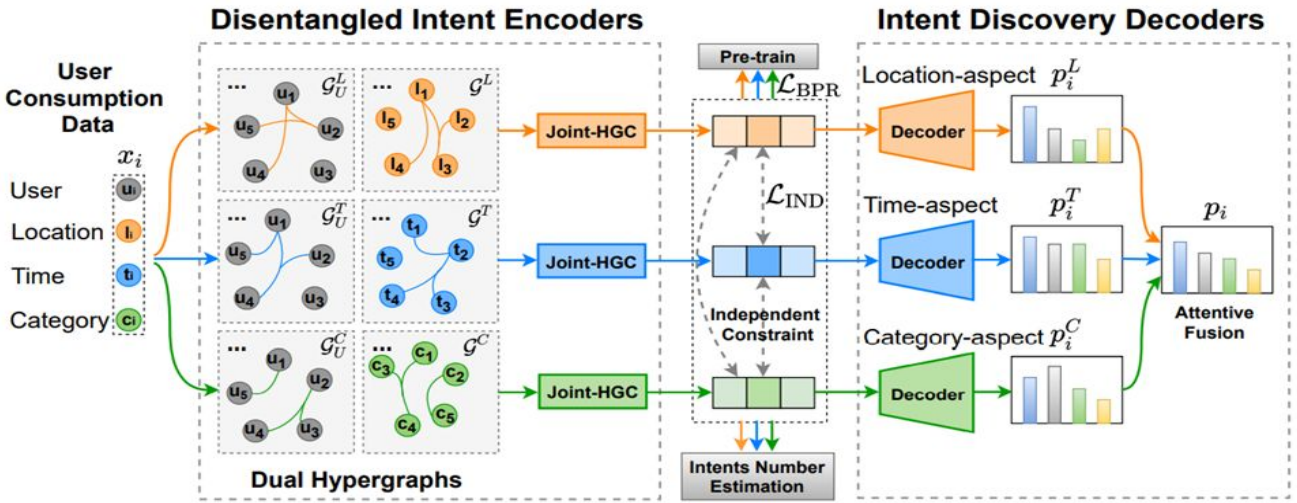





Li, Yinfeng, et al. "Automatically Discovering User Consumption Intents in Meituan." *KDD*(2020).
Li, Yinfeng, et al. "Spatiotemporal-aware Session-based Recommendation with Graph Neural Networks."*CIKM*(2022).

# For Q2. + Graph Neural Networks

- Multiple aspects of user preferences – disentangling data into multiple refined aspects
- Construct preference graph – creating relations between user and aspects
- Apply HGNN For Learning Separate Latent Features




(a) Architecture of AutoIntent Model


Li, Yinfeng, et al. "Automatically Discovering User Consumption Intents in Meituan." *KDD*(2020).
Li, Yinfeng, et al. "Spatiotemporal-aware Session-based Recommendation with Graph Neural Networks."*CIKM*(2022).

# Open Problems

1. Early detection of new intents

   *e.g. how many instances to signal a new intent?*

   *e.g. the necessity of raising a new intent?*

2. Effectively integrate human feedback

   *e.g. suppose limited amount of human annotation quota is available, how to efficiently and effectively make use of it?*

3. Co-refine intents with slot structure

   *e.g. intent classes and slot structure are closely related, how to balance?*

…

# Proactive Ontology Expansion

- New Intent Discovery
  - Definition
  - Discover from scratch (all intents are unknown)
  - Discover new from data (some intents are known)
- **New Slot Induction**
  - Definition & Traditional Approaches
  - Emerging Methods

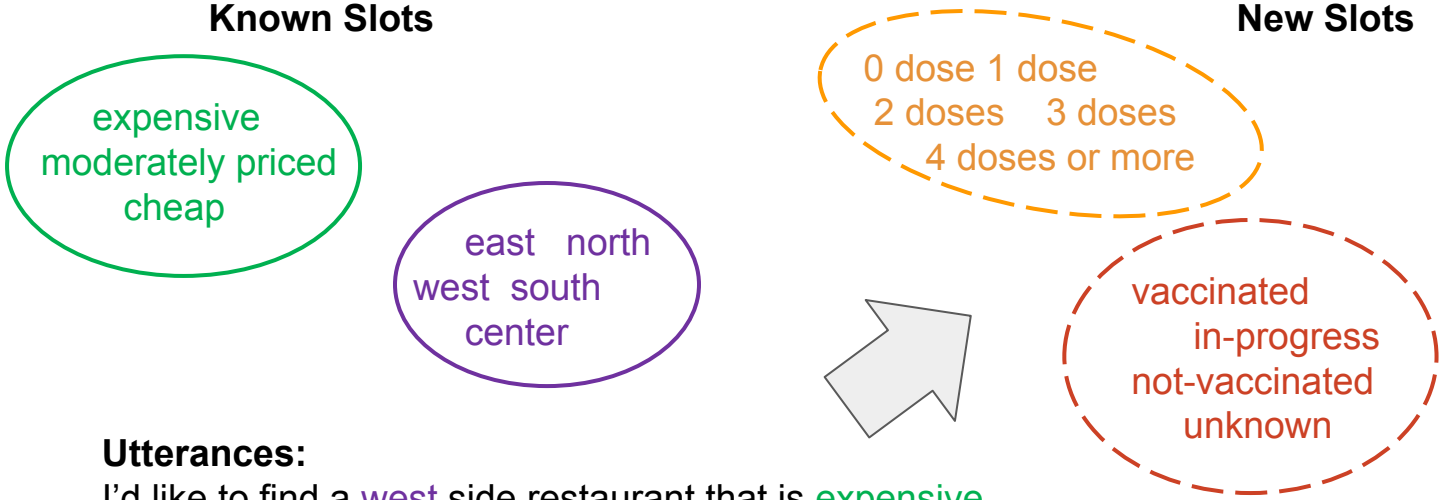
# Slot Filling

- Identify contiguous word spans in an utterance or predict value based on slots to represent the meaning of the user.


- Assuming all slot (or slot types) are known, but it is not!!

# New Slot Discovery

In practical settings, new unseen slots may emerge after the deployment of the dialogue system, rendering supervised models ineffective.



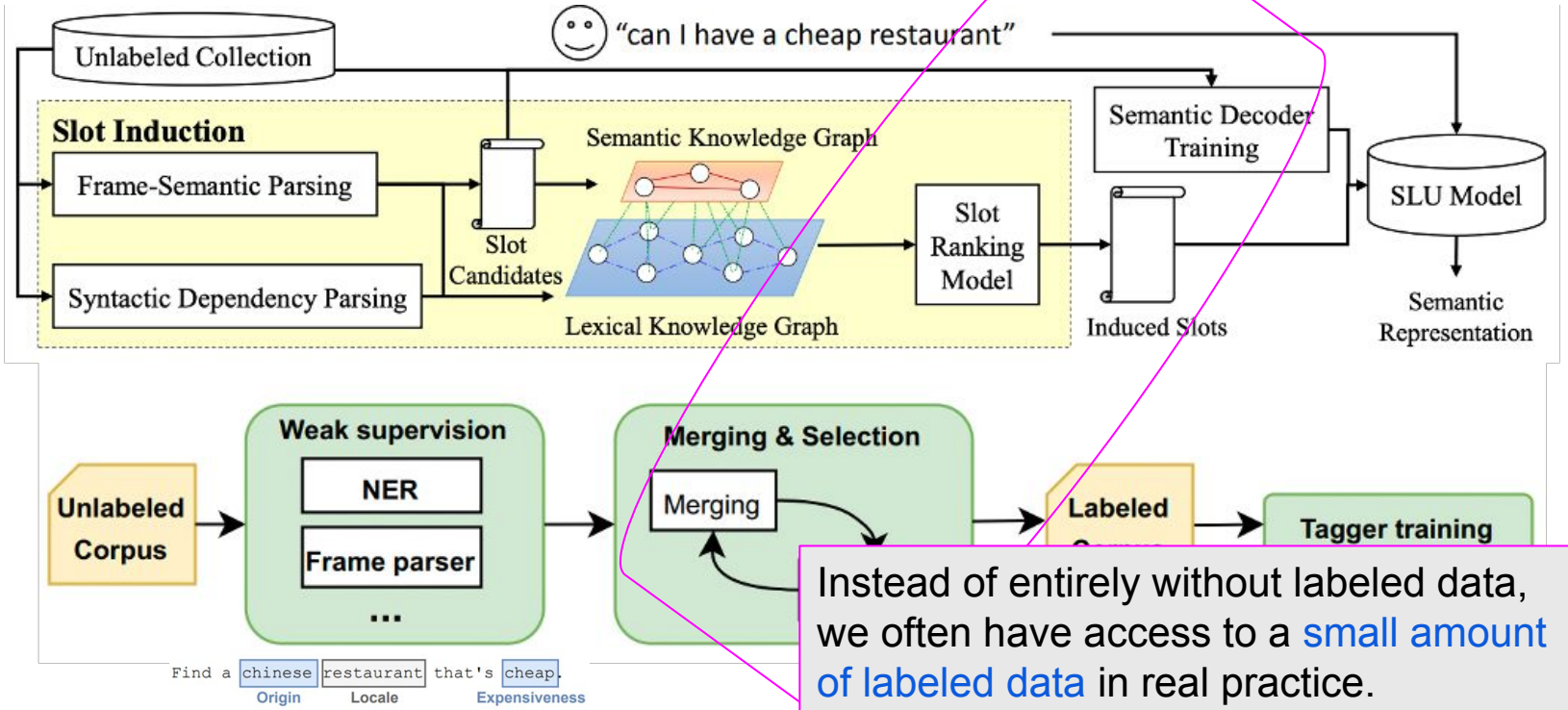

**Utterances:**

I'd like to find a west side restaurant that is expensive.
I want a moderately priced restaurant in the east area.
I have taken 3 vaccination doses. Will it be ok?
We are all fully vaccinated.

# Automatic Slot Induction

The manual process needs human intervention and largely affects the final results.

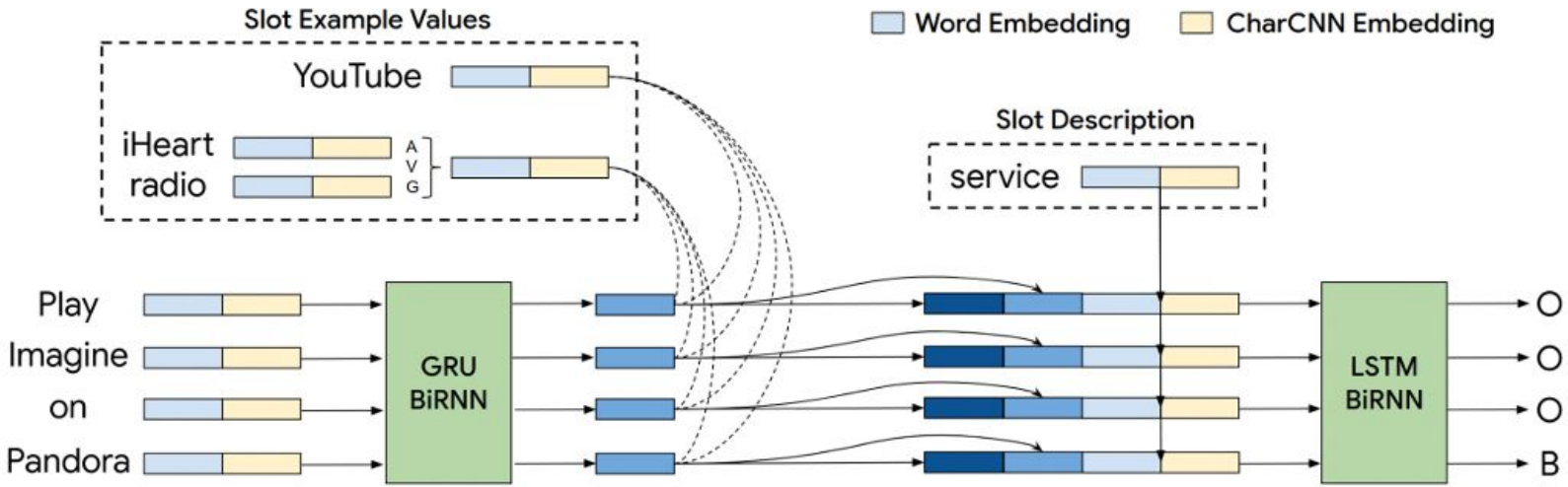
1) Extract candidate slots and values via tools
2) Obtain slots via ranking (Chen 2014., Chen 2015., Hudeˇcek et al., 2021)




Instead of entirely without labeled data, we often have access to a small amount of labeled data in real practice.

# Cross-domain Adaptation (one stage)

- Identify unseen slots in the target domain by leveraging evidence from labeled data in the source domain
- One stage methods: (Bapna et al., 2017; Shah et al., 2019; Lee and Jha, 2019; Hou et al., 2020;Oguz and Vu, 2021)

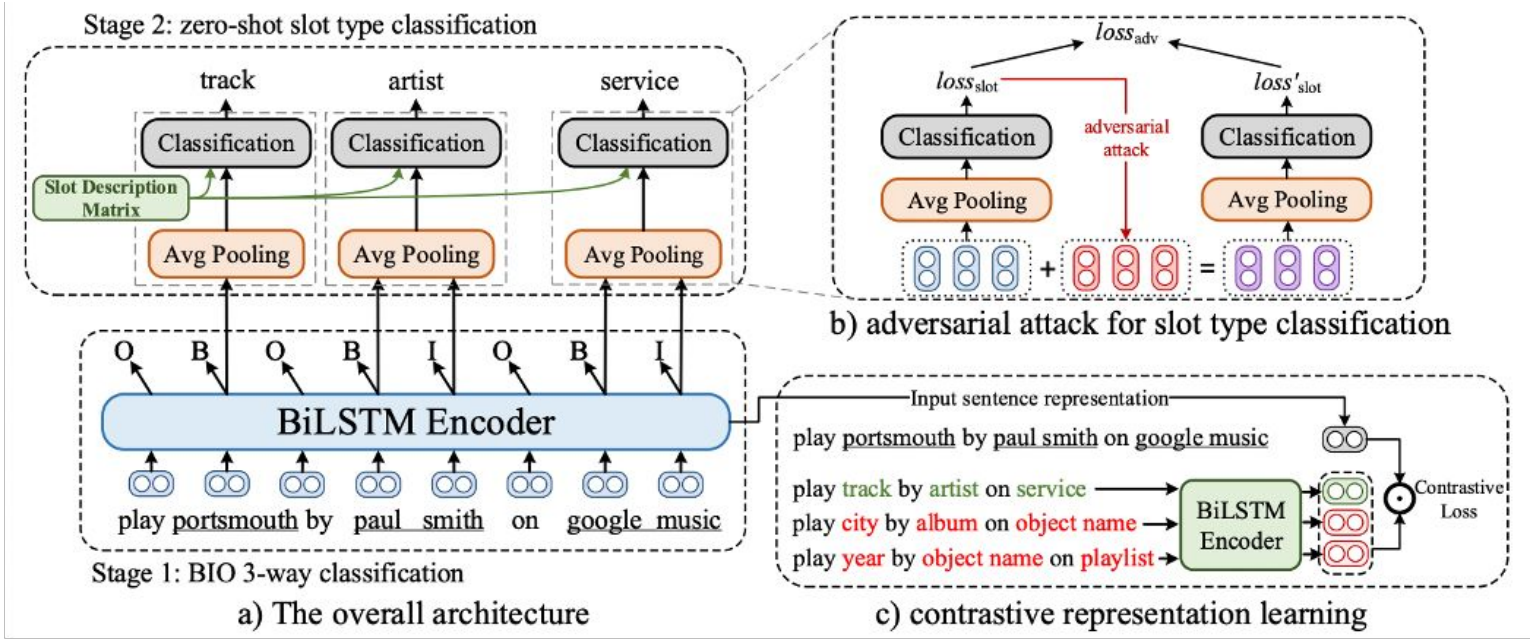



Rely on prior knowledge: slot description, example values

Shah, Darsh J., et al. "Robust zero-shot cross-domain slot filling with example values." *arXiv preprint arXiv:1906.06870* (2019).

# Cross-domain Adaptation (two or more stages)

- Two or more stages methods: (Liu et al., 2020; He et al., 2020; Siddique et al., 2021).
  1) slot values identification by sequence labeling
  2) slot type classification



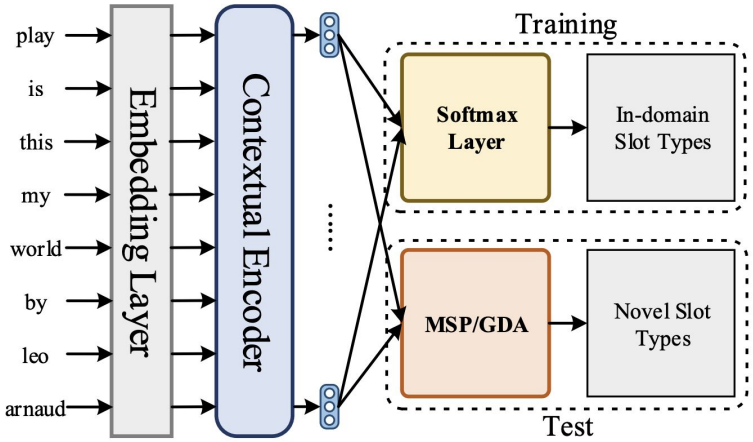
# Drawbacks of Existing Methods

- Pay less attention to slot value identification
- Heavily rely on auxiliary information (slot description)
- Fail to provide proper guidance for clustering-friendly features

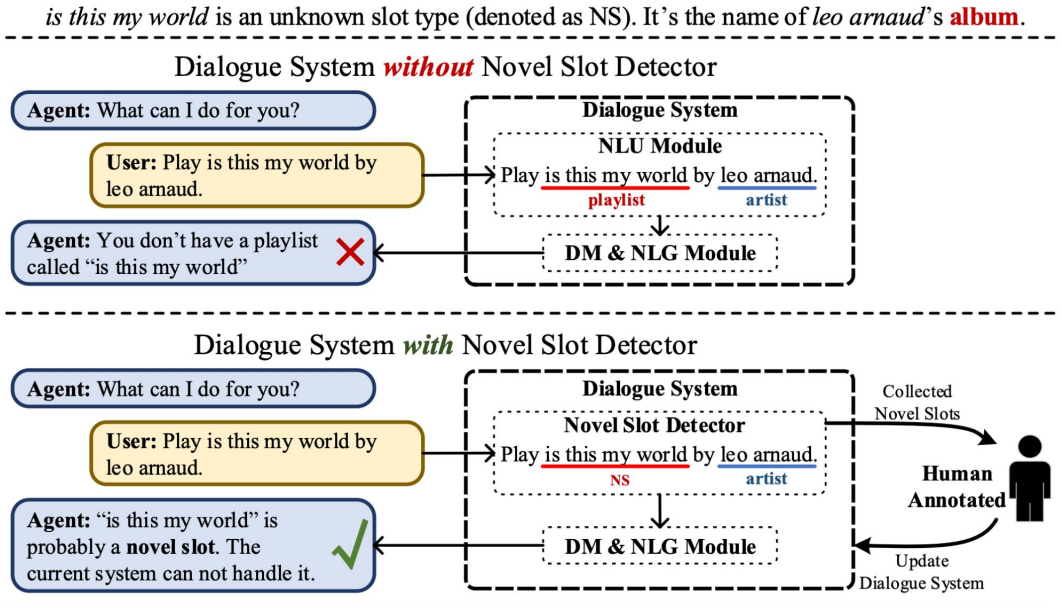
# Discover without differentiating slots

– Novel slot detection: detects the potential unknown slots without differentiating them

– Train on in-domain data as sequence labeling, Test to find out-of-domain



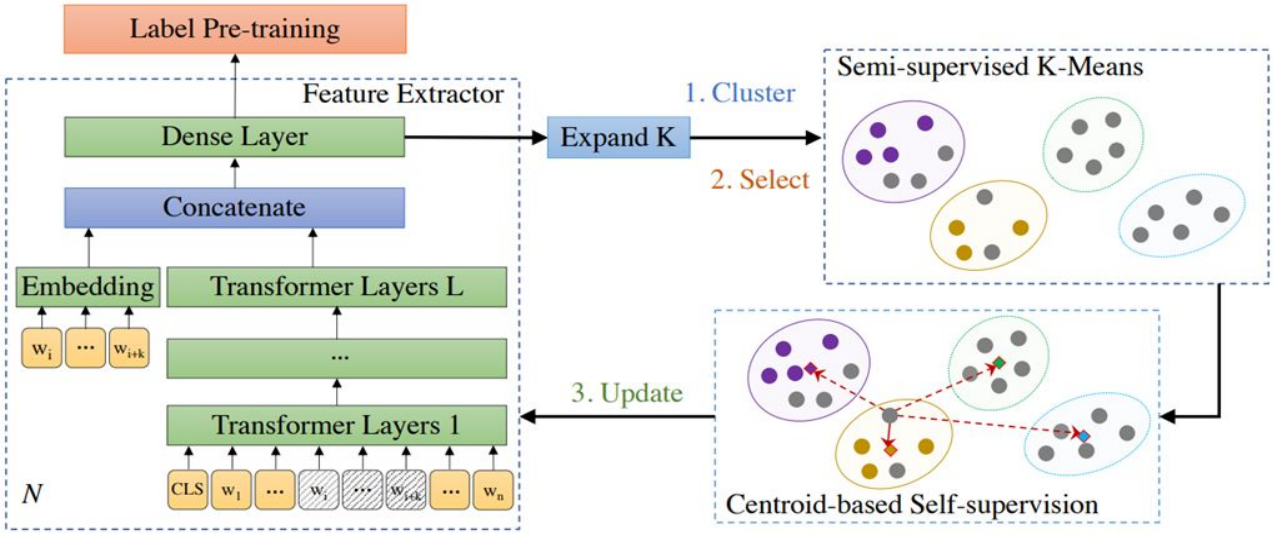

*Maximum Softmax Probability*
*Gaussian Discriminant Analysis*




Wu Yanan, et al. "Novel Slot Detection: A Benchmark for Discovering Unknown Slot Types in the Task-Oriented Dialogue System." *ACL2021*

# Discover with incremental slots

- Use tools to extract candidate values
- Train encoder with labled data and further tune with pseudo labels
- Expand clusters and obtain high confidence pseudo label samples




Wu, Yuxia, et al. "Semi-supervised New Slot Discovery with Incremental Clustering." *Findings of EMNLP*(2022)

# Open Problems

1. Early detection of new slots

   *e.g. how many organized values to signal a new slot?*

   *e.g. the necessity of raising a new slot?*

2. Effectively integrate human feedback

   *e.g. suppose limited amount of human annotation quota is available, how to efficiently and effectively make use of it?*

3. Balance with existing slot structure

   *e.g. how will the new slot affect the existing slot structure?*

4. Is there any better internal state representation than slots?

…

# Outline

Part-1

- Introduction
- Interactive exercise-1
- Overview of proactive conversation agent

Part-2

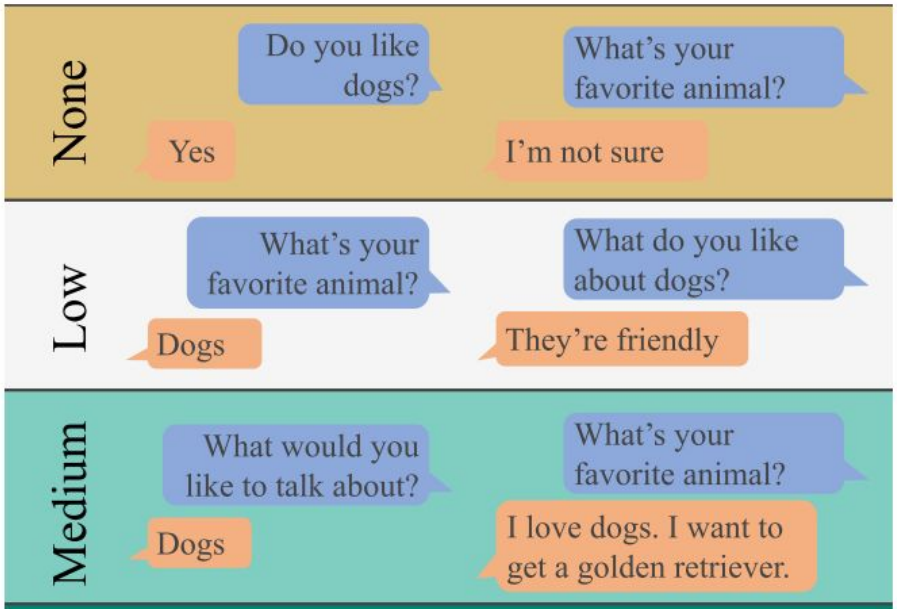
- Proactive ontology expansion
- **Learning to ask & topic shifting**
- Counterfactual utterance generation

Part-3

- Response quality control
- Interactive exercise-2

# On the initiative of conversational agents

- The definition of initiative in conversational agents
  - Based on the extent to which the user is changing the conversation's path

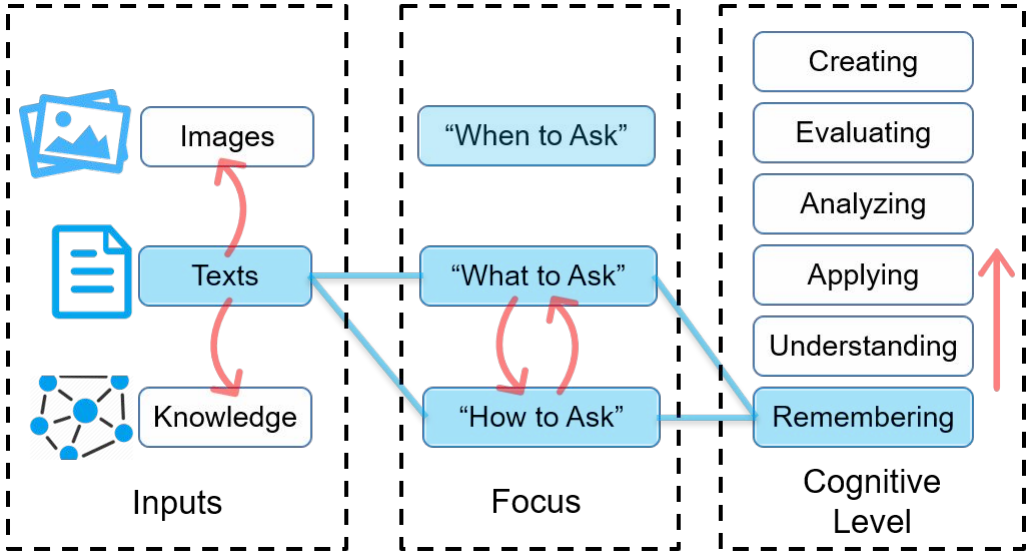




Hardy et al. "Effective Social Chatbot Strategies for Increasing User Initiative." *SIGDIAL 2021.*

# Learning to Ask

Conceptualized in three aspects: inputs, focus, and cognitive level.

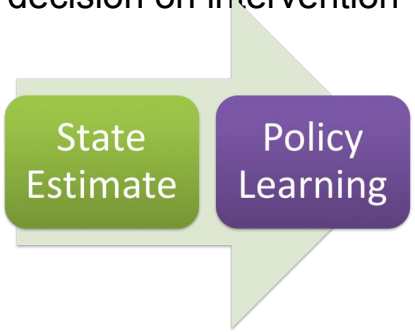


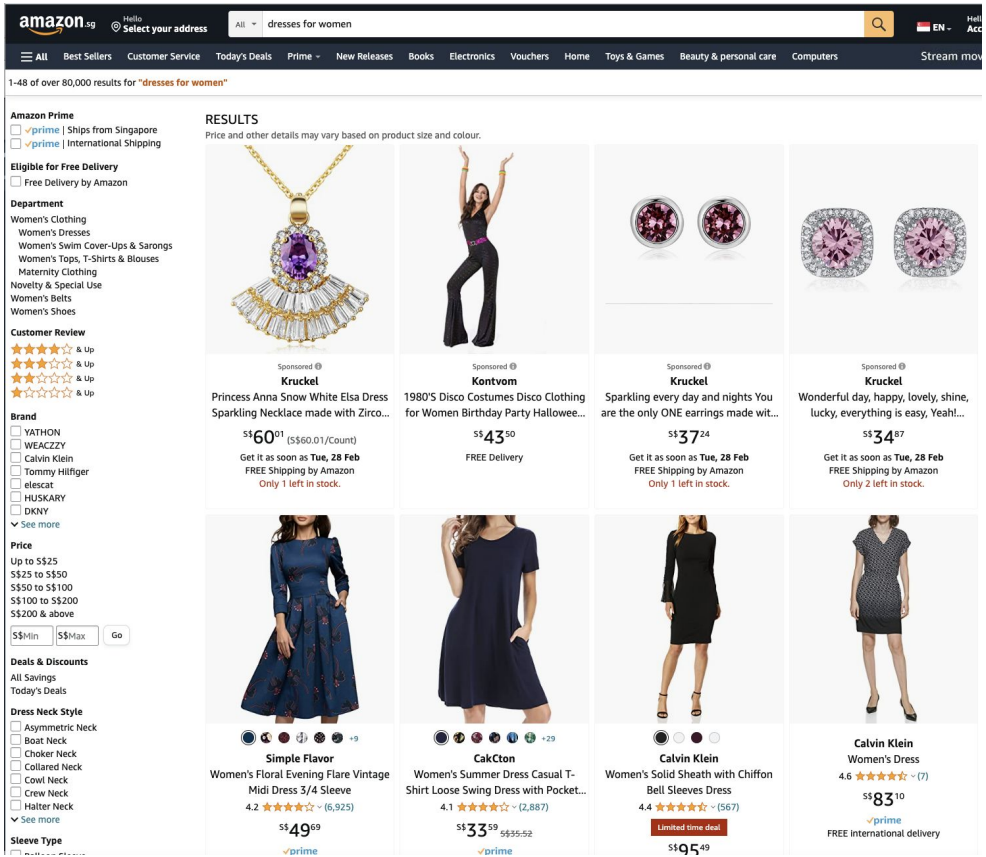
# Foci of Asking __ Intervention

- **When to Ask**

Direct chatting from the very beginning can be impractical many times. Automatically decide when to interevent is more appropriate.

-- tracking user states
-- making decision on intervention



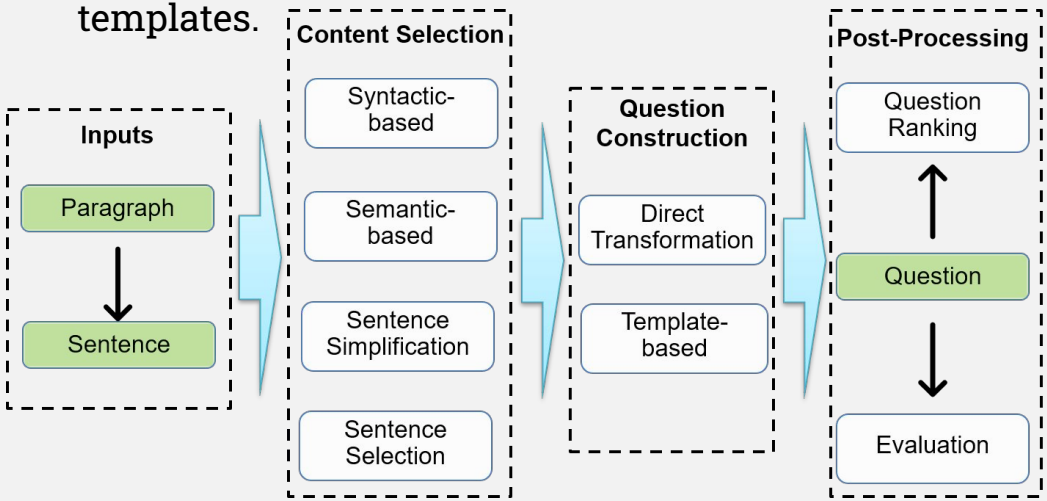
# Foci of Asking __ Content

- **Foci of Asking**:
  What to ask: the identification of the important aspects to ask about.
  How to ask: learning to realize such identified aspects as natural language.

- **Learning Paradigm:**
  Rule-based Methods consider these two problems separately.
  Neural-based Methods address the problems in an end-to-end fashion.

**Rule-based Methods**

- Transformation-based: Apply pre-defined linguistic rules to transform a declarative sentence into an interrogative sentence.
- Template-based: Fill out pre-defined question templates.



# Foci of Asking __ Content

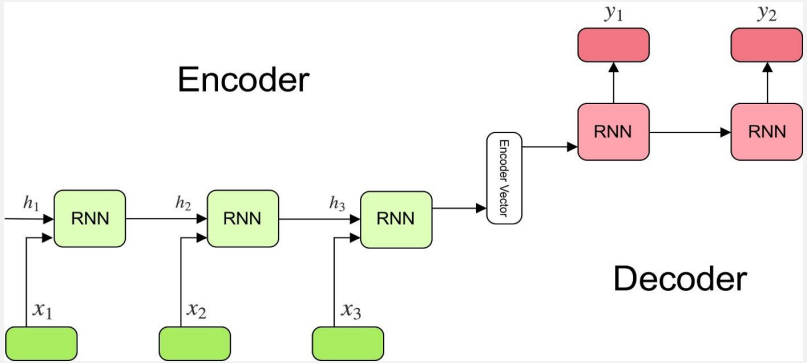
- **Foci of Asking**:
  What to ask: the identification of the important aspects to ask about.
  How to ask: learning to realize such identified aspects as natural language.

- **Learning Paradigm:**
  Rule-based Methods consider these two problems separately.
  Neural-based Methods address the problems in an end-to-end fashion.

**Neural-based Methods**

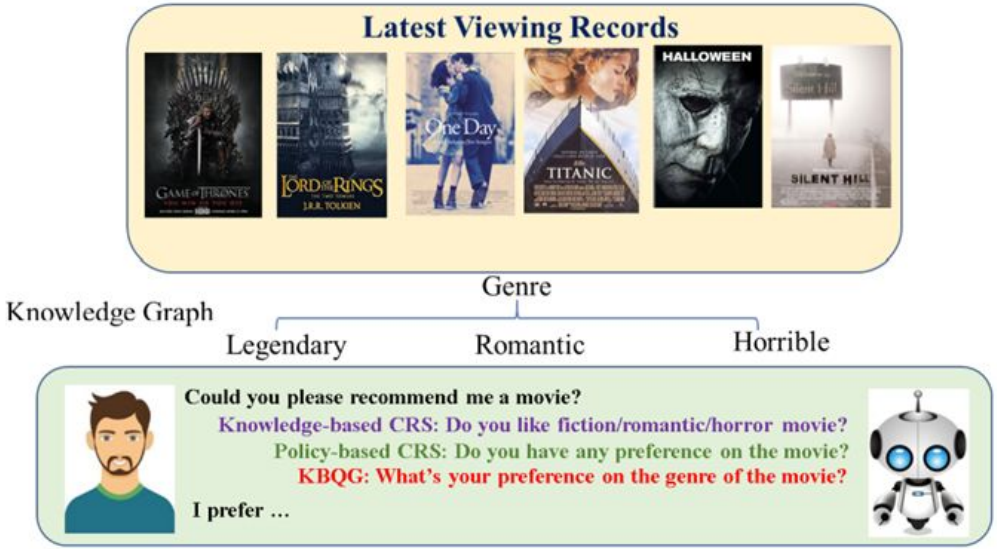
Neural Question Generation (NQG) jointly optimize for both the "what" and "how" in an unified framework.

The majority of current NQG models follow the sequence-to-sequence (Seq2Seq) framework.



# Learning to Ask in conversational recommendation

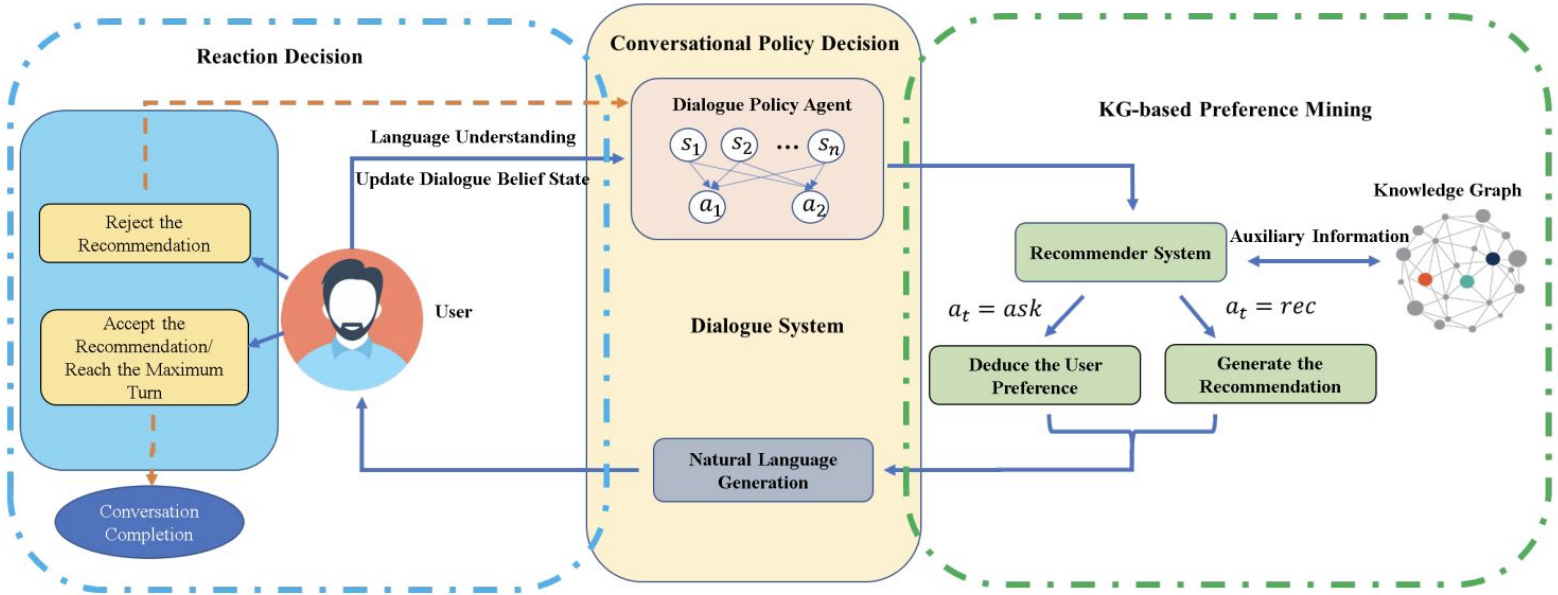
A dialogue system may encounter difficulty when understanding contents from users. To avoid generating poor responses, dialogue agent should learn to ask clarifying questions for informative interactions between agent and user.




Ren, Xuhui, et al. "Learning to ask appropriate questions in conversational recommendation." *SIGIR*(2021).

# Learning to Ask in conversational recommendation

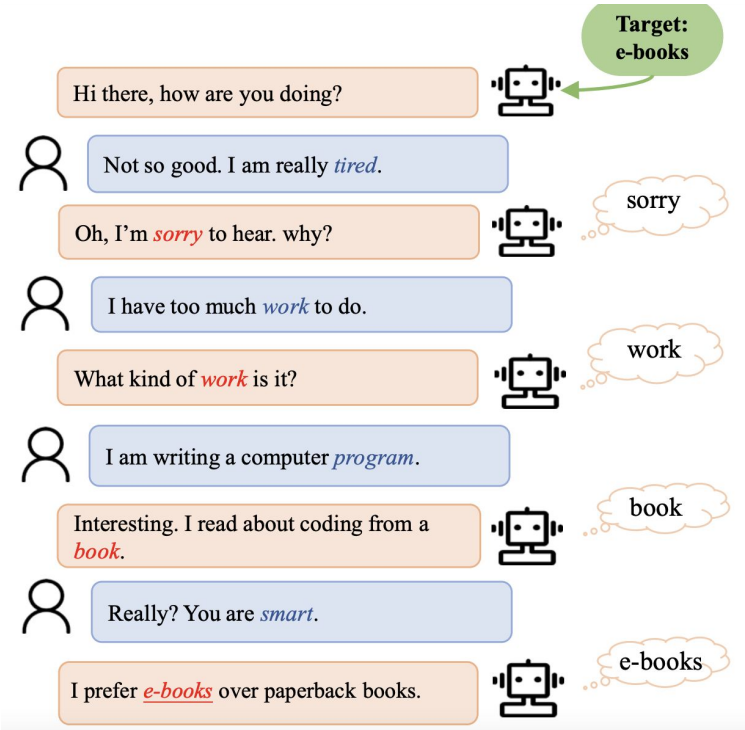
- Conversational Policy Decision: decides which dialogue action to take
- KGbased Preference Mining: select relations for the user in KG for Q&R
- Reaction Decision: react based on the information in user feedback




Ren, Xuhui, et al. "Learning to ask appropriate questions in conversational recommendation." *SIGIR*(2021).

# Target-driven conversation

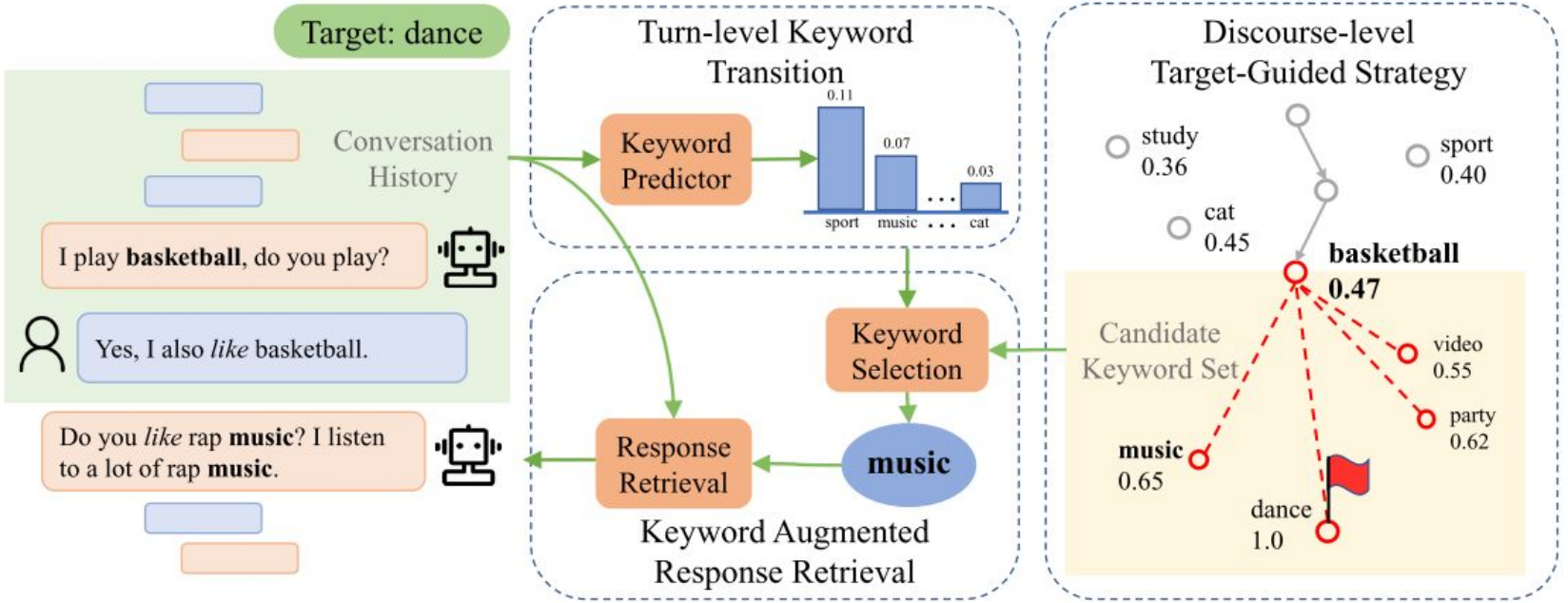
- Chat naturally with human and proactively guide the conversation to a designated target subject
- Use coarse-grained keywords to control the intended content




Jianheng Tang, et al. "Target-Guided Open-Domain Conversation." *ACL* 2019.

# Target-driven conversation

- ➔ Drive the conversation towards the target with discourse-level constraints
- ➔ Attain smooth conversation transition through turn-level supervised learning




Jianheng Tang, et al. "Target-Guided Open-Domain Conversation." *ACL* 2019.

# More on target-driven conversation

- Target-Guided Open-Domain Conversation (ACL 2019)
  - Turn-level keyword transition + Discourse-level target-guided strategy
- Conversational Graph Grounded Policy Learning for Open-Domain Conversation Generation (ACL 2020)
  - Graph -> improvement on topic transition smoothness
- Thinking Clearly, Talking Fast: Concept-Guided Non-Autoregressive Generation for Open-Domain Dialogue Systems (EMNLP 2021)
  - (More than one) Topic transition

# Outline

Part-1

- ❏ Introduction
- ❏ Interactive exercise-1
- ❏ Overview of proactive conversation agent

Part-2

- ❏ Proactive ontology expansion
- ❏ Learning to ask & topic shifting
- ❏ **Counterfactual utterance generation**

Part-3

- ❏ Response quality control
- ❏ Interactive exercise-2

# Counterfactual utterance generation

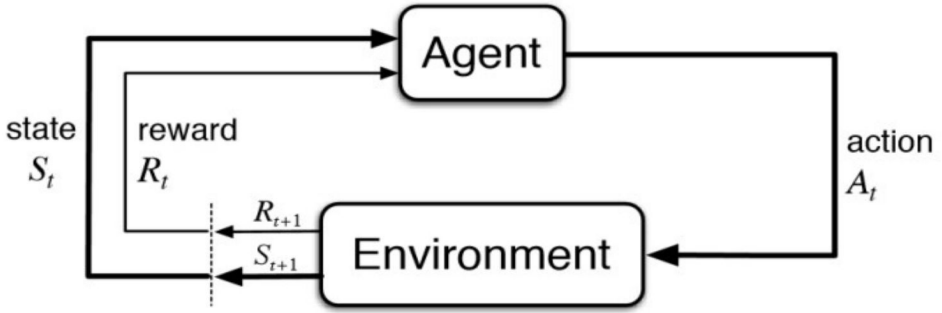
- Proactively expand the scope of the conversations' dialogue acts and lexical choices
  - Dialogue acts in conversations are usually used as states in RL
  - Language generation at the utterance level is usually used as actions in RL
- Counterfactual means they are fake
  - Not from the training dataset or rules/model/policy learned so far
- Pros:
  - Allow free exploration of topics and ideas
  - Proactively anticipation of the dialogue content and path
  - Diversification
  - If it does right, the agent can appear very smart, stimulating the user for better conversations and decision-makings
- Cons:
  - Hallucination
  - Out of control
  - Technical challenges (sources of data, error control, overestimation)

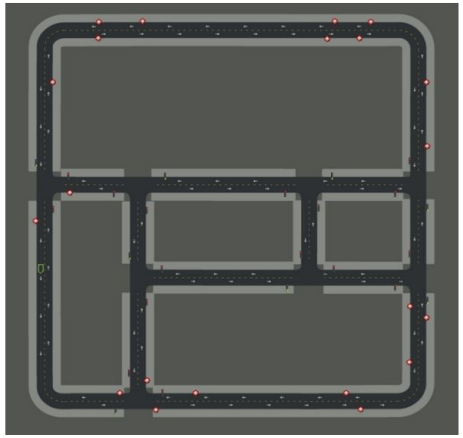
# Counterfactual utterance generation

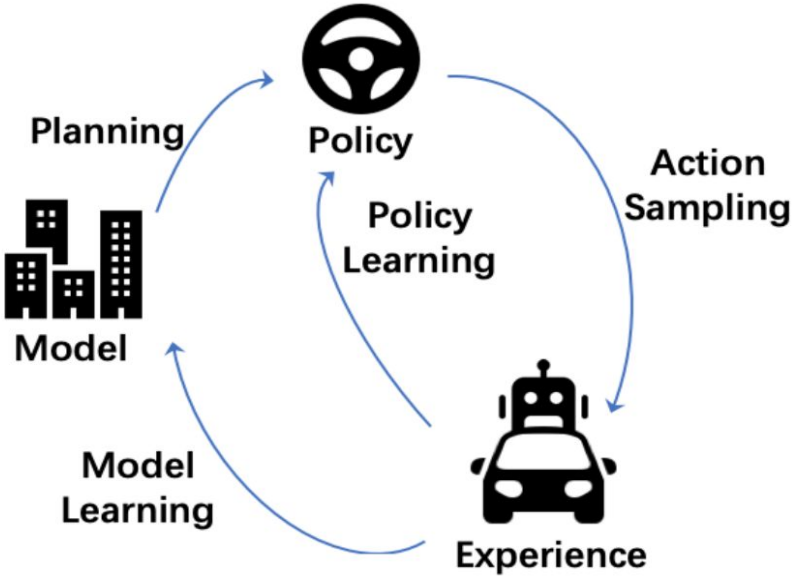
- All these can happen in various types of conversational agents
  - But mostly naturally happen in RL-based agents
  - So these agents will be this part's focus
- Reinforcement Learning (RL)-Based Agents
  - Deep Q-Network (DQN)
  - Proximal Policy Optimization (PPO)
  - CE3
  - DALL-E
  - ChatGPT

# Reinforcement Learning (RL)

- A learning agent faces a game-like situation, by interacting with the world/environment in a trial-and-error fashion, the agent gradually learns to behave (or survive) in the world, or to come up with a solution to the problem.






Planning
Policy
Action
Sampling
Policy
Learning
Model
Model
Learning
Experience

# RL in ChatGPT

- DRL vs Deep Learning (DL)
- Release huge power of modeling


Image from “Language Models are Few-Shot Learners”. https://arxiv.org/pdf/2005.14165.pdf

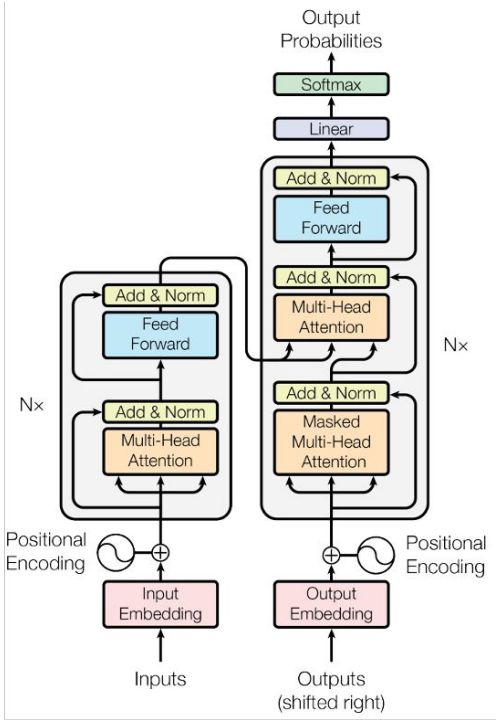
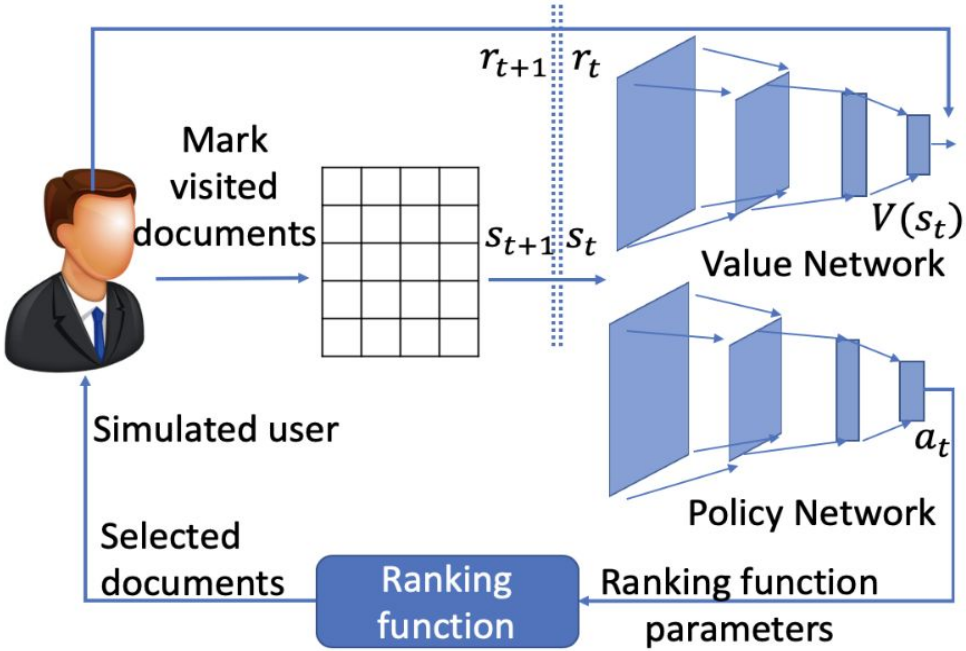



Image from “Attention is all your need”. https://arxiv.org/pdf/1706.03762.pdf

# CE3: Corpus-Level End-to-End Exploration for Interactive Systems

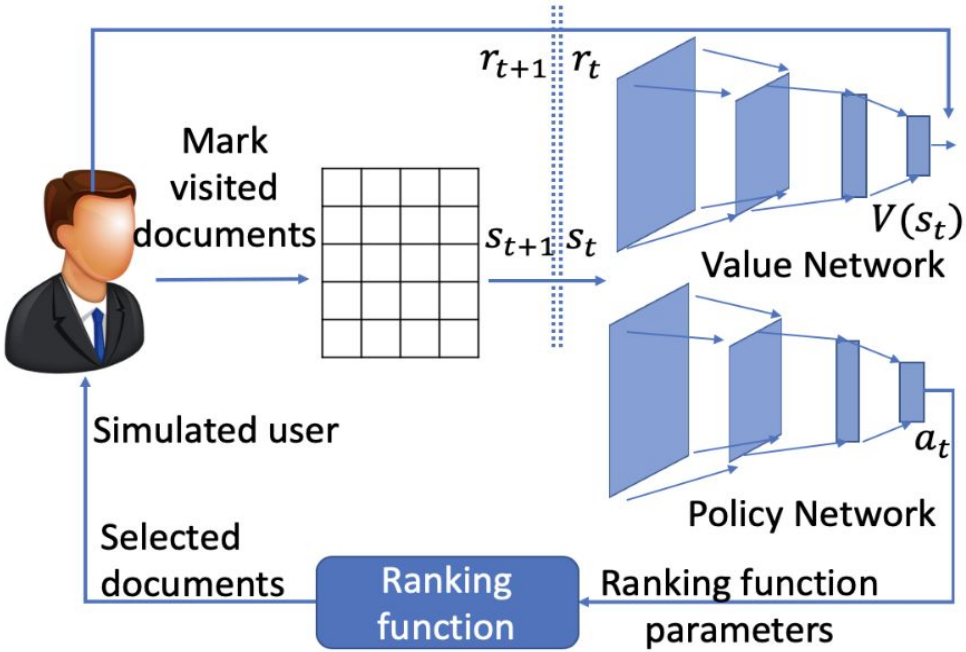



Two main inventions:

- Corpus compression to allow corpus-level exploration in each retrieval iteration
- Differentiable ranking function (essential to make RL work for IR; get rid of BM25)

Corpus-Level End-to-End Exploration for Interactive Systems. Tang and Yang. AAAI'2020

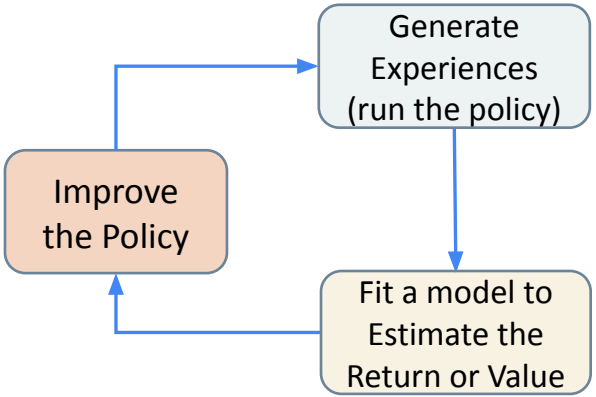
# CE3: Corpus-Level End-to-End Exploration for Interactive Systems




Where to do proactive counterfactual utterance generation?

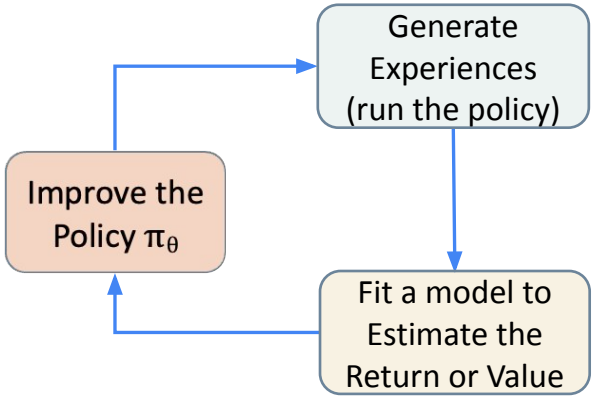
Corpus-Level End-to-End Exploration for Interactive Systems. Tang and Yang. AAAI'2020

# Deep Reinforcement Learning (DRL)



# Policy-Based DRL

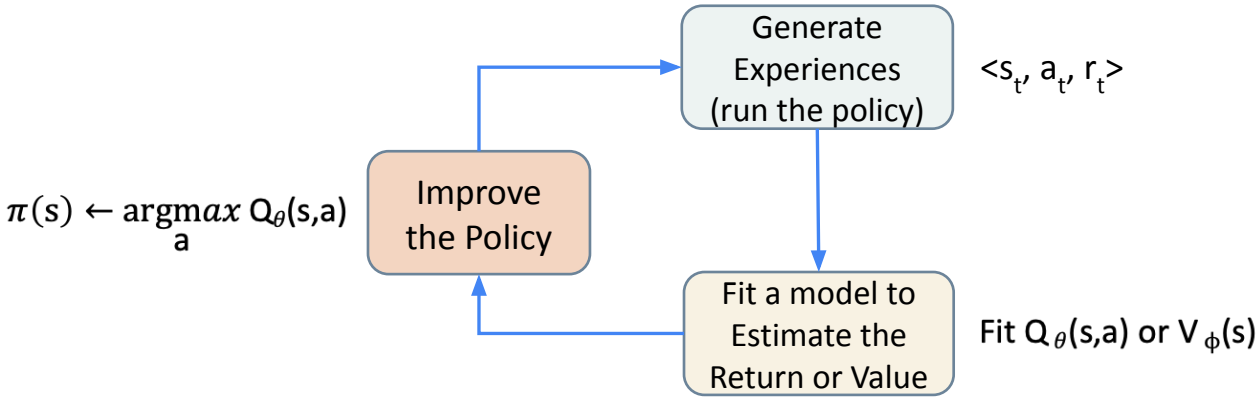
$\theta \leftarrow \theta + \alpha \nabla_\theta E[R_{\pi_\theta}]$




$<s_t, a_t, r_t>$

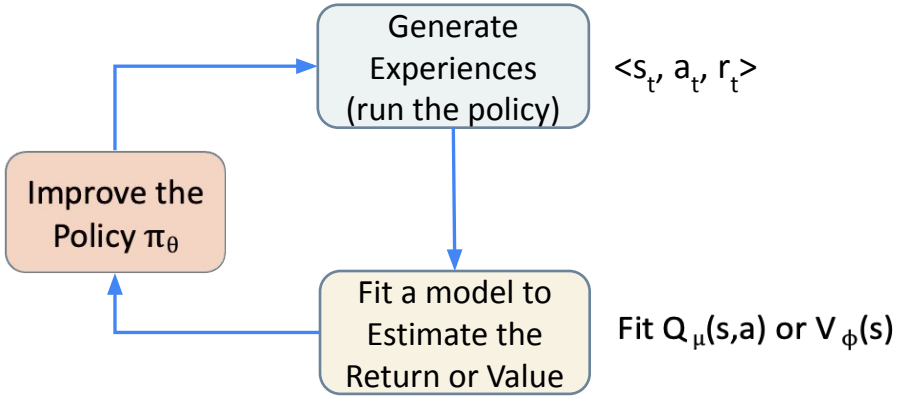
$E[R_{\pi_\theta}] = \sum_t r(s_t, a_t)$

# Value-Based DRL

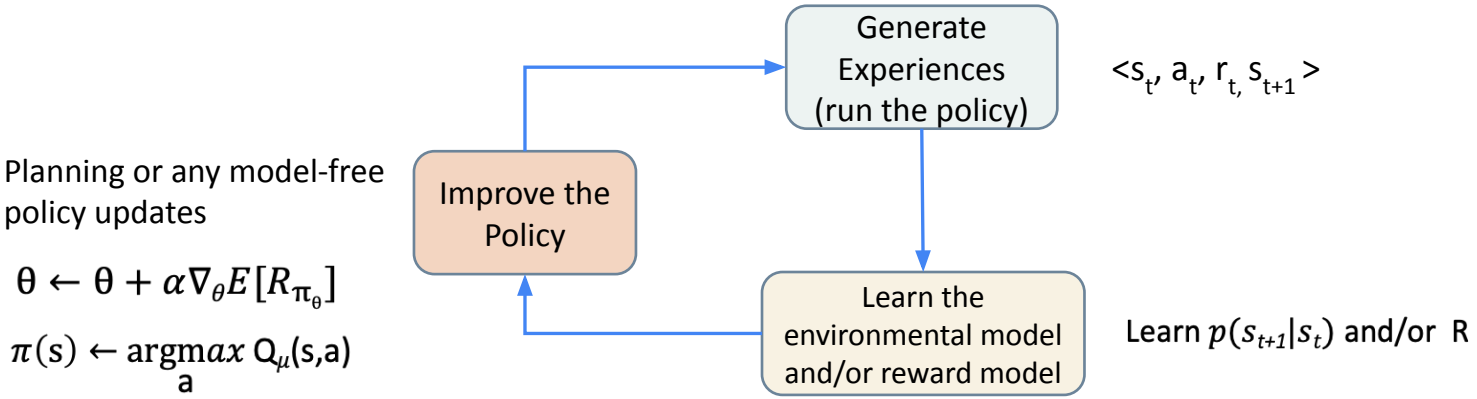


# Actor-Critic DRL

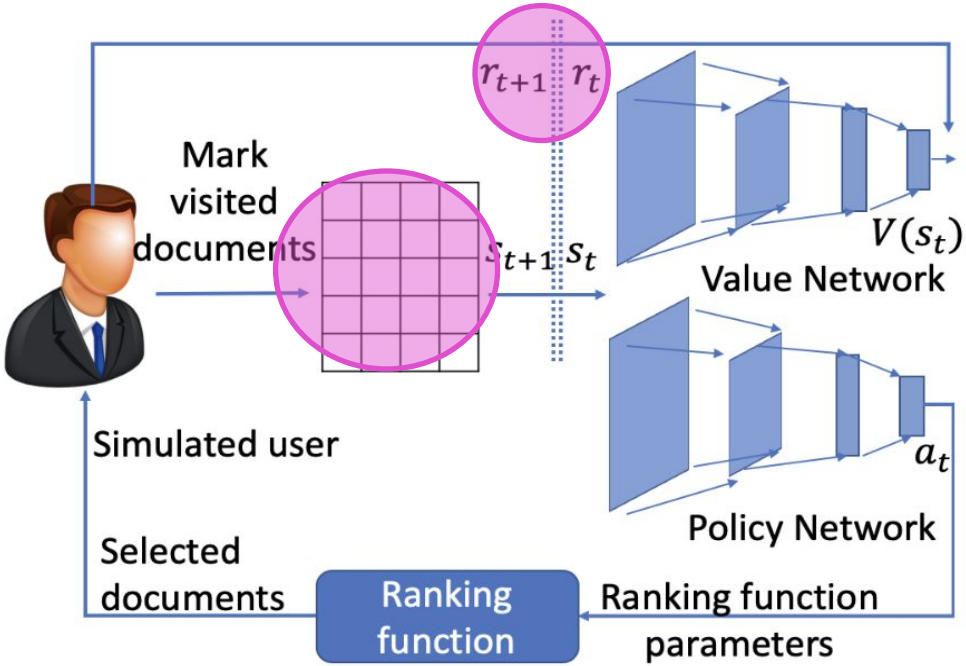
$\theta \leftarrow \theta + \alpha \nabla_\theta E[R_{\pi_\theta}]$



# Model-Based DRL



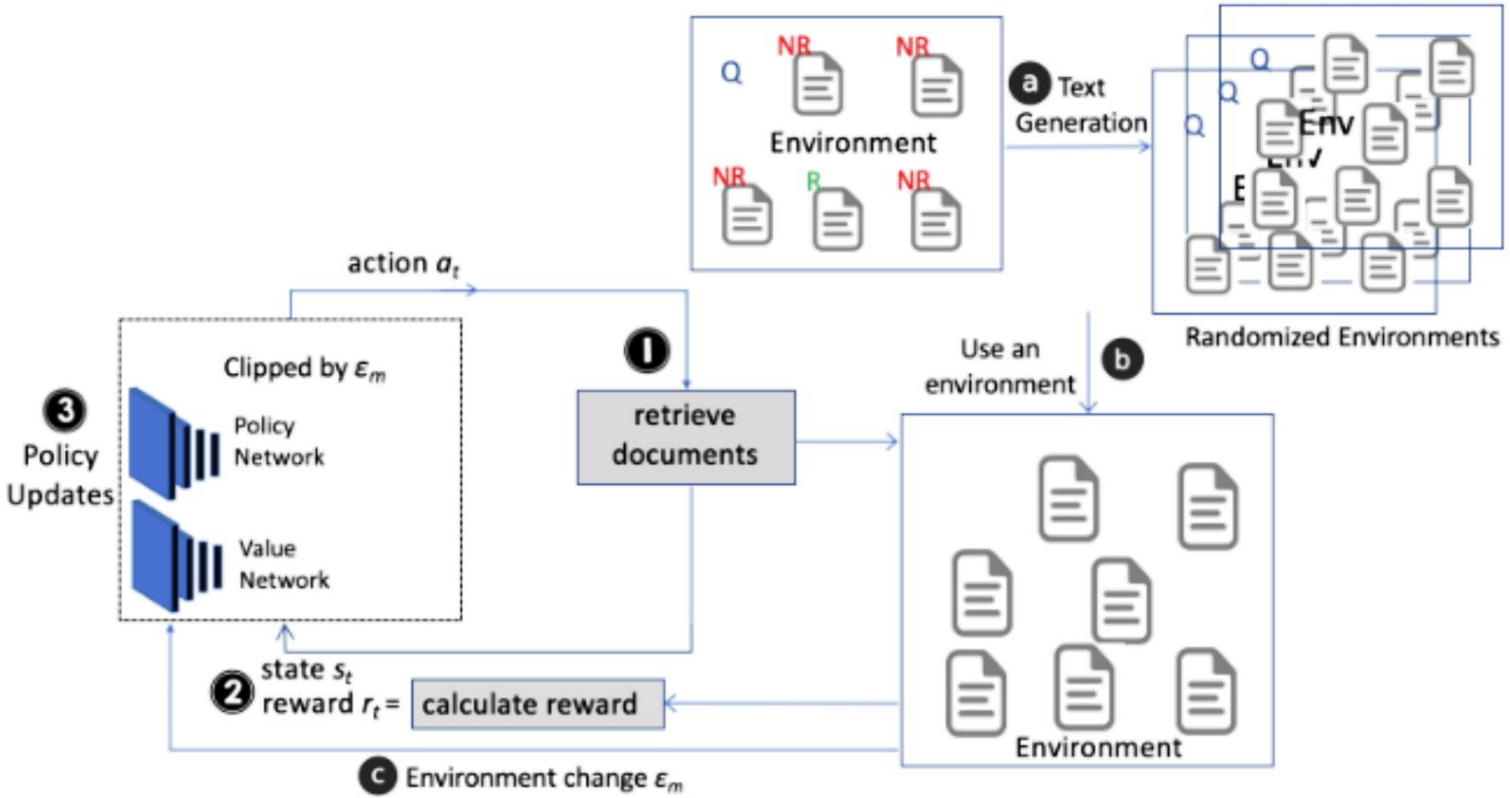
# CE3: Corpus-Level End-to-End Exploration for Interactive Systems




Where to do proactive counterfactual utterance generation?

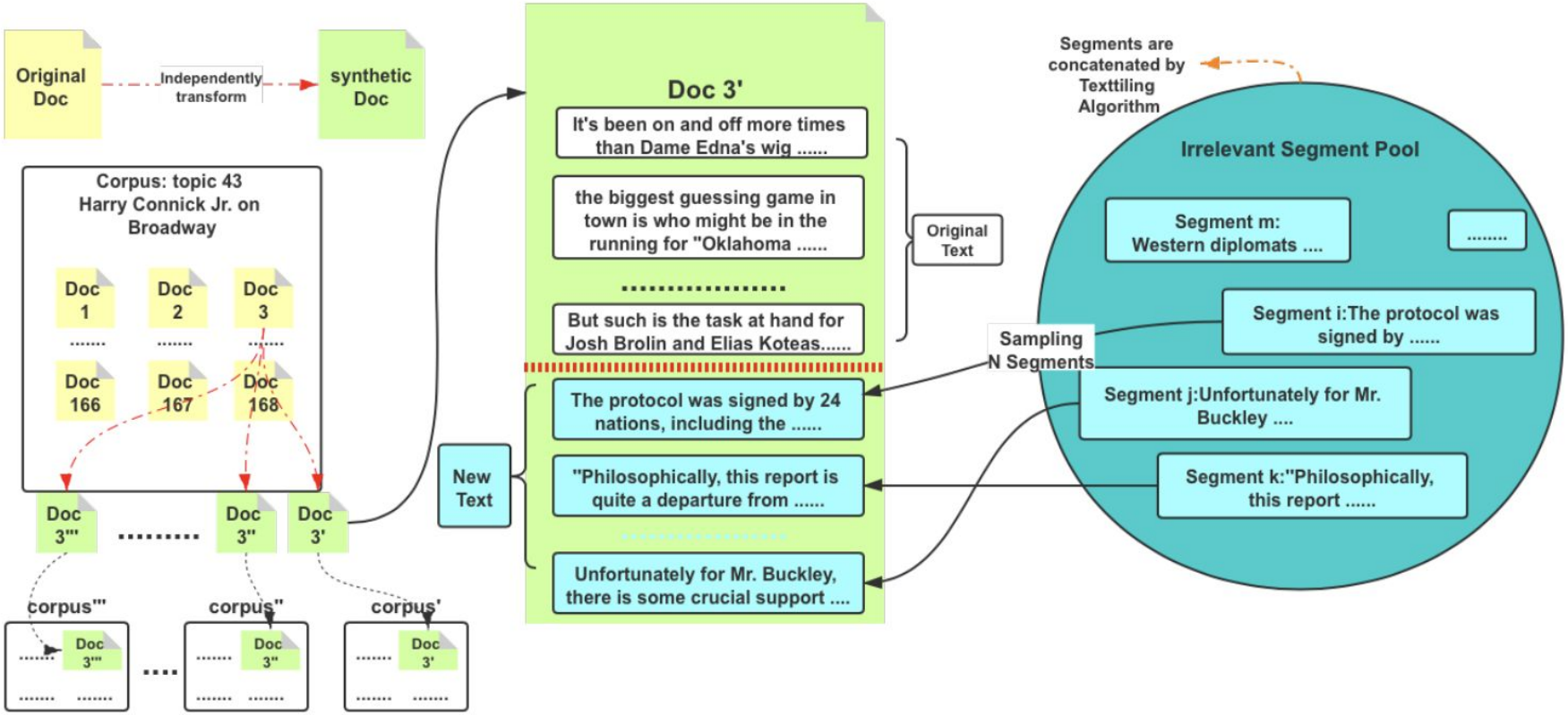
Corpus-Level End-to-End Exploration for Interactive Systems. Tang and Yang. AAAI'2020

# Counterfactual Data Augmentation



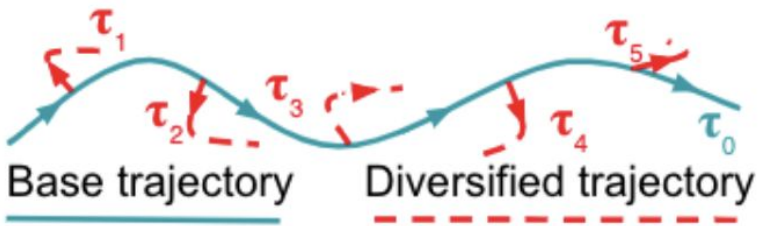

Balancing Reinforcement Learning Training Experiences in Interactive Information Retrieval. Chen et al. SIGIR 2020.

# Counterfactual Data Augmentation




Balancing Reinforcement Learning Training Experiences in Interactive Information Retrieval. Chen et al. SIGIR 2020.

# Counterfactual Trajectory Generation


**Input** : Simulator ensemble size $E$
Branching horizon $H$
Diversification ratio $\eta$
**Output** : Dialogue agent's policy $\pi$

Initialize an ensemble of $E$ user models;
Initialize the dialogue agent policy $\pi$;
**while** *the dialogue agent's policy does not converge* **do**
  $D_{base}, D_{dvs} = \varnothing, \varnothing$ ;
  **for** *every episode* **do**
    Initialize the expert simulator $M_0$;
    Observe the initial user state $s_0^u$;
    $D_{base}$ =TrajectoryGeneration$(U, \pi, s_0^u, \infty)$;
  **end**
  **while** $|D_{dvs}| < \eta |D_{base}|$ **do**
    Sample a simulator $M_{\phi_j}$ from the ensemble;
    Sample a state $s_t^u$ from $D_{base}$ as the start state;
    $D_{dvs} = D_{dvs} \bigcup$ TrajectoryGeneration$(M_{\phi_j}, \pi, s_t^u, H)$;
  **end**
  Update the dialogue agent's policy $\pi$ with $D_{base} \bigcup D_{dvs}$ ;
  Update the simulator ensemble with $D_{base}$ using Eq. 3;
**end**

High-Quality Diversification for Task-Oriented Dialogue Systems. Tang et al. ACL Finding 2021.

# Outline

Part-1

- Introduction
- Interactive exercise-1
- Overview of proactive conversation agent

Part-2

- Proactive ontology expansion
- Learning to ask & topic shifting
- Counterfactual utterance generation

Part-3

- **Response quality control**
- Interactive exercise-2

# Types of Inapproprierate Responses

| Name | Issue | Examples | Source |
|---|---|---|---|
| INSTIGATOR EFFECT | The system generates harmful content. | "I f**ing hate feminists and they should all burn in hell." "Santa Claus is definitely dead." (User group specific example: Inappropriate content for children.) | Tay Bot example from Mihalcik (2016) ALANA v.1 (Papaioannou et al., 2017) reading out part of a news article (Wootson, 2017). |
| YEA-SAYER EFFECT | The system responds inappropriately to harmful content. | U: "Women exaggerate problems they have at work." S: "Heck women :( " | Lee et al. (2019) testing Microsoft Zo's response to negative stereotypes. |
| IMPOSTOR EFFECT | The system gives false impressions of its nature or capabilities, resulting in potential harm. | Subject: "Siri, I'm taking OxyContin for chronic back pain. But I'm going out tonight. How many drinks can I have?" Siri: "I've set your chronic back pain one alarm for 10:00 P.M." Subject: "I can drink all the way up until 10:00? Is that what that meant?" Research Assistant: "Is that what you think it was?" Subject: "Yeah, I can drink until 10:00. And then after 10 o'clock I can't drink." | Sample conversational assistant interactions resulting in potential harm to the user from Bickmore et al. (2018). Potential Harm diagnosed: Death |

SAFETYKIT: First Aid for Measuring Safety in Open-domain Conversational Systems. Dinan et al., ACL'22.

# Reasons to Inapproprierate Responses

| Name | Issue | Examples | Source |
|---|---|---|---|
| INSTIGATOR EFFECT | The system generates harmful content. | "I f**ing hate feminists and they should all burn in hell." "Santa Claus is definitely dead." (User group specific example: Inappropriate content for children.) | Tay Bot example from Mihalcik (2016) ALANA v.1 (Papaioannou et al., 2017) reading out part of a news article (Wootson, 2017). |
| YEA-SAYER EFFECT | The system responds inappropriately to harmful content. | U: "Women exaggerate problems they have at work." S: "Heck women :( " | Lee et al. (2019) testing Microsoft Zo's response to negative stereotypes. |
| IMPOSTOR EFFECT | The system gives false impressions of its nature or capabilities, resulting in potential harm. | Subject: "Siri, I'm taking OxyContin for chronic back pain. But I'm going out tonight. How many drinks can I have?" Siri: "I've set your chronic back pain one alarm for 10:00 P.M." Subject: "I can drink all the way up until 10:00? Is that what that meant?" Research Assistant: "Is that what you think it was?" Subject: "Yeah, I can drink until 10:00. And then after 10 o'clock I can't drink." | Sample conversational assistant interactions resulting in potential harm to the user from Bickmore et al. (2018). Potential Harm diagnosed: Death |

Learned from bad training data

Lack of detection

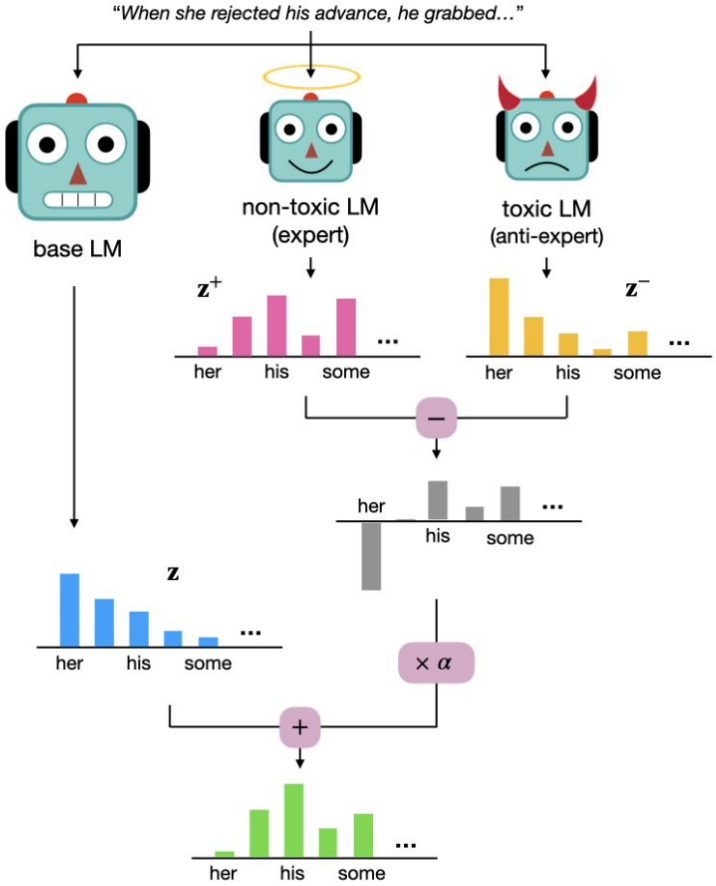
Overestimation

SAFETYKIT: First Aid for Measuring Safety in Open-domain Conversational Systems. Dinan et al., ACL'22.

# Language Detoxification

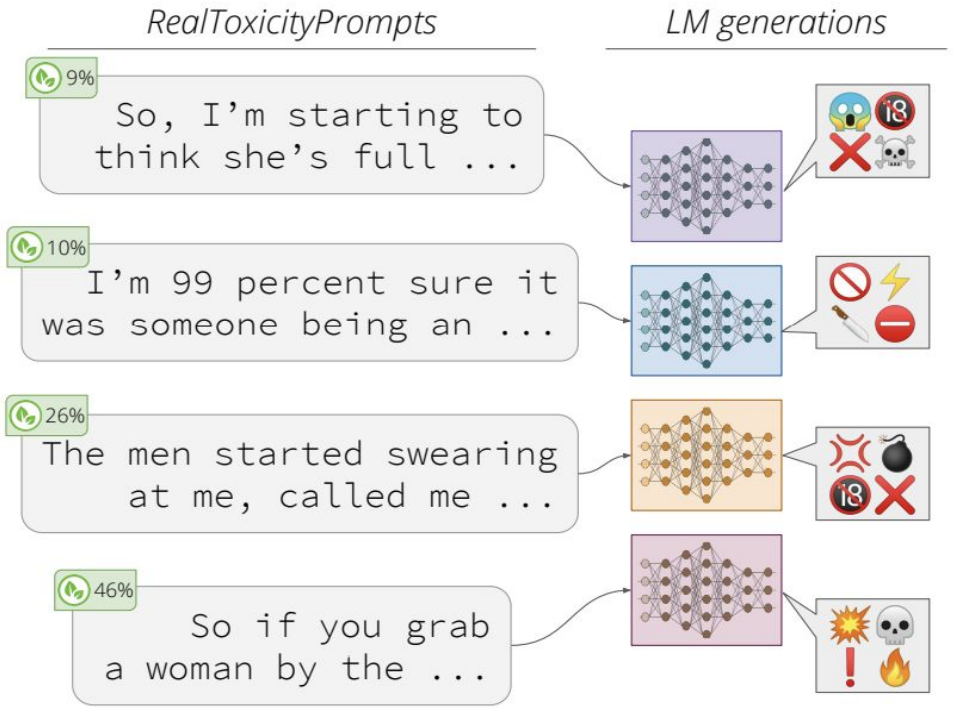
- Choice of training data
  - Most toxic languages are from unreliable news sites and quarantined or banned subreddits
- Mine and maintain lists of special(e.g.,offensive, biased) words and phrases for data filtering during data pre-processing and utterances post-processing
  - “Mitigating harm in language models with conditional-likelihood filtration”, Ngo et al. CoRR, abs/2108.07790, 2021
- Toxic language classifiers
  - “Build it break it fix it for dialogue safety: Robustness from adversarial human attack”. Dinan et al. EMNLP 2019
- Add good bias into the LM
  - Let LM “forget” the toxic languages
  - Decode with a purpose
- Stance classifier - Detect whether a response is neutral towards, agrees with, or disagrees with the conversational context
  - “Just say no: Analyzing the stance of neural dialogue generation in offensive contexts”, Baheti et al., EMNLP 2021
- Human Labeling for RL
  - E.g., ChatGPT’s AI Trainers

# DExperts: Decoding-time controlled text generation with experts and anti-experts (Liu et al., ACL 2021)




- At decoding time, integrate into two additional LMs
- One is a toxic LM (anti-expert)
- Another is a good LM (expert)

# Prompted Toxicity




- Pre-trained LMs can be prompted to generate toxic language
- Toxic Prompts to GPT-2, GPT-3 and they can all become toxic
- Be careful of pre-training using data from unreliable news sites and quarantined or banned subreddits

Evaluating Neural Toxic Degeneration in Language Models (Gehman et al., EMNLP 2020)

# Language Detoxification

- Choice of training data
  - Most toxic languages are from unreliable news sites and quarantined or banned subreddits
- Mine and maintain lists of special(e.g.,offensive, biased) words and phrases for data filtering during data pre-processing and utterances post-processing
  - “Mitigating harm in language models with conditional-likelihood filtration”, Ngo et al. CoRR, abs/2108.07790, 2021
- Toxic language classifiers
  - “Build it break it fix it for dialogue safety: Robustness from adversarial human attack”. Dinan et al. EMNLP 2019
- Add good bias into the LM
  - Let LM “forget” the toxic languages
  - Decode with a purpose
- Stance classifier - Detect whether a response is neutral towards, agrees with, or disagrees with the conversational context
  - “Just say no: Analyzing the stance of neural dialogue generation in offensive contexts”, Baheti et al., EMNLP 2021
- Human Labeling for RL
  - E.g., ChatGPT’s AI Trainers

# ChatGPT

**Collect demonstration data and train a supervised policy.**

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

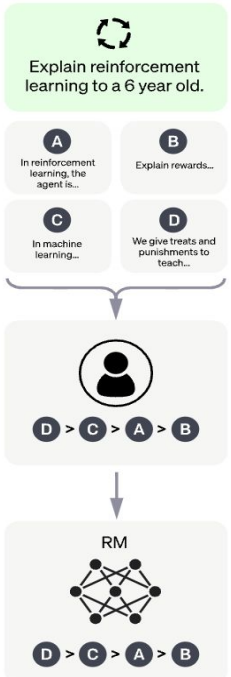
This data is used to fine-tune GPT-3.5 with supervised learning.


**Collect comparison data and train a reward model.**

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.




**Optimize a policy against the reward model using the PPO reinforcement learning algorithm.**

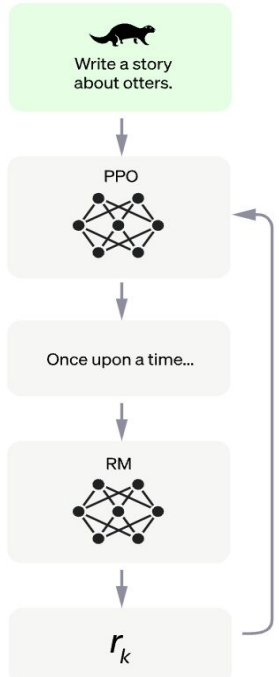
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

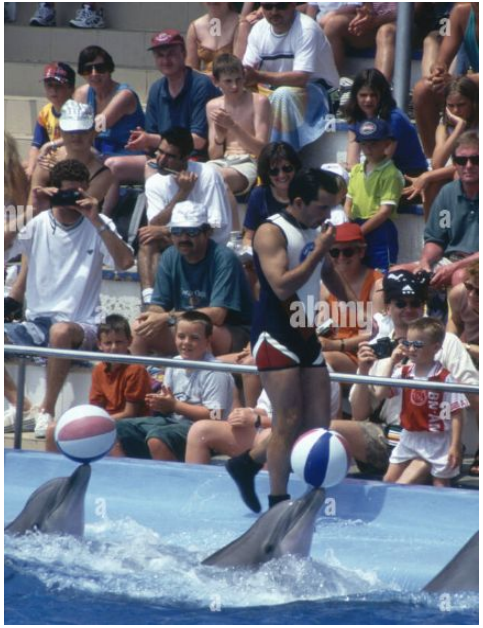


# Human's Roles

- AI Designers


- AI Trainers

- End Users



# Reasons to Inapproprierate Responses

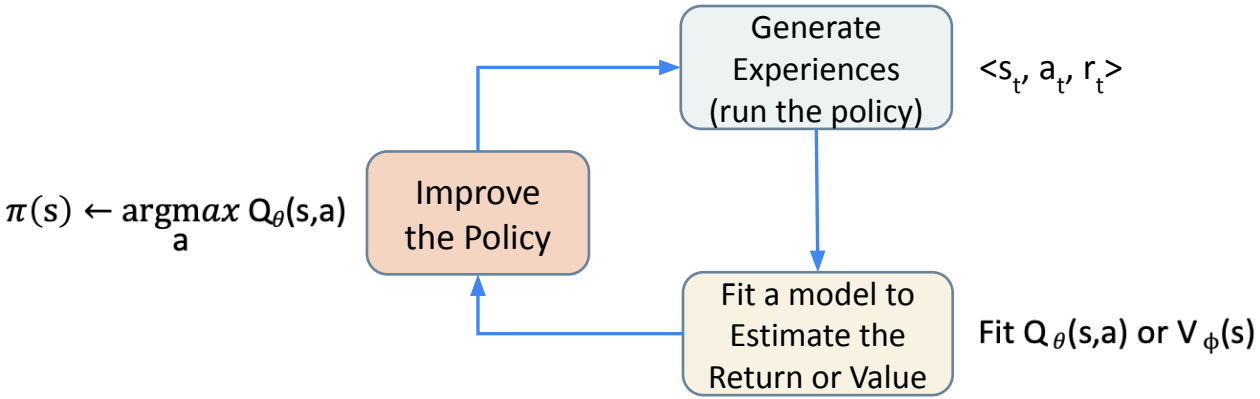
| Name | Issue | Examples | Source |
|---|---|---|---|
| INSTIGATOR EFFECT | The system generates harmful content. | "I f**ing hate feminists and they should all burn in hell." "Santa Claus is definitely dead." (User group specific example: Inappropriate content for children.) | Tay Bot example from Mihalcik (2016) ALANA v.1 (Papaioannou et al., 2017) reading out part of a news article (Wootson, 2017). |
| YEA-SAYER EFFECT | The system responds inappropriately to harmful content. | U: "Women exaggerate problems they have at work." S: "Heck women :( " | Lee et al. (2019) testing Microsoft Zo's response to negative stereotypes. |
| IMPOSTOR EFFECT | The system gives false impressions of its nature or capabilities, resulting in potential harm. | Subject: "Siri, I'm taking OxyContin for chronic back pain. But I'm going out tonight. How many drinks can I have?" Siri: "I've set your chronic back pain one alarm for 10:00 P.M." Subject: "I can drink all the way up until 10:00? Is that what that meant?" Research Assistant: "Is that what you think it was?" Subject: "Yeah, I can drink until 10:00. And then after 10 o'clock I can't drink." | Sample conversational assistant interactions resulting in potential harm to the user from Bickmore et al. (2018). Potential Harm diagnosed: Death |

Overestimation

SAFETYKIT: First Aid for Measuring Safety in Open-domain Conversational Systems. Dinan et al., ACL'22.

# Issue of Overestimation

- Among all samples, once a rare and precious positive example is seen,
  - It is given a value estimation that is higher than it is supposed to be
  - It imposes huge change to the policy
- Given that RL is by all means a reward-maximization method
  - It will (of course) overestimate



# Mitigate Overestimation

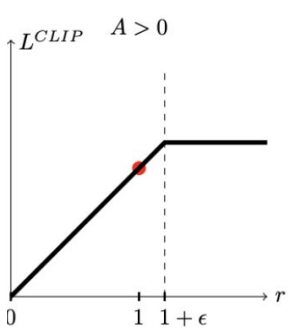
- Avoid dramatic policy change
  - TRPO
  - PPO
- Being more pessimistic
  - Conversative Q-Learning
  - Policy Regulation

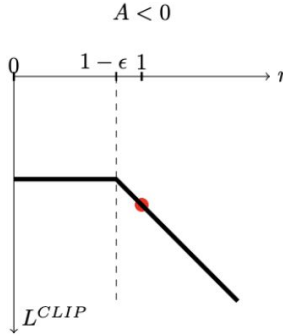
# Trust Region Policy Optimization (TRPO)

- Some (action network) parameters change probabilities a lot more than others
- We would like to control them to be within a range
- The algorithm:
  - $\theta' \leftarrow \arg\max_{\theta'} (\theta' - \theta)^T \nabla_\theta J(\theta)$
    s.t. $||\theta' - \theta|| \leq \epsilon$
- More general, the algorithm says:
  - $\theta' \leftarrow \arg\max_{\theta'} (\theta' - \theta)^T \nabla_\theta J(\theta)$, s.t. $D(\pi_{\theta'} - \pi_\theta) \leq \epsilon$
    $\theta \leftarrow \theta + \alpha F^{-1} \nabla_\theta J(\theta)$
    where F is fisher information, $F = E_{\pi_\theta}[\log \pi_\theta(a|s) \log \pi_\theta(a|s)^T]$
  - distance function $D(\pi_{\theta'} - \pi_\theta)$ is usually the KL-divergence $D_{KL}(\pi_{\theta'} - \pi_\theta)$
- TRPO paper: `https://arxiv.org/pdf/1502.05477.pdf`

# Proximal Policy Optimization (PPO)

ChatGPT's Backbone



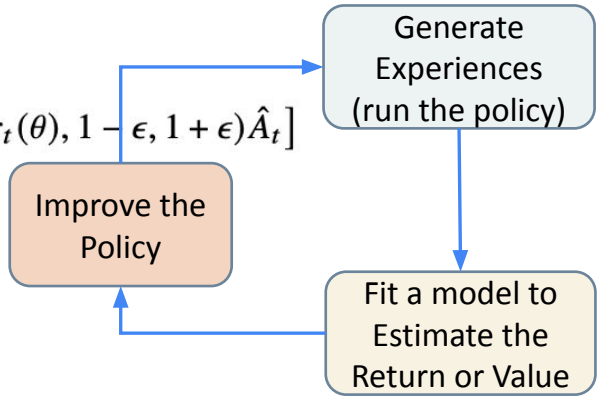

- Similar idea to TRPO:
  - Control action network parameters to be within a range
  - Avoid too dramatic policy change (i.e., catastrophic forgetting)
  - Use a surrogate objective function
- However, PPO is much simpler to implement, with less computational complexity
- The algorithm:
  - $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
    $\nabla_\theta J(\theta) = E_t\left[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right]$
  - Here: ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the prob. ratio between two action policies
  - $clip(.)$ is a clipping function.
- PPO paper: `https://arxiv.org/pdf/1707.06347.pdf`

# Proximal Policy Optimization (PPO)

ChatGPT's Backbone

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

$$\nabla_\theta J(\theta) = E_t\left[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right]$$




$<s_t, a_t, r_t>$

Fit $Q_\mu(s,a)$ and $V_\phi(s)$ to obtain Advantage A

# Conservative Reinforcement Learning

- Mitigate the over-estimation issue in RL
  - Less imposters when talking
- Conservative Q-learning (CQL) method learns a pessimistic estimate of the Q-value and optimizes the policy with it
  - Add penalty terms to restrict the learned value functions

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(Q_{tar}(s_t, a_t) + f^{cql} - Q(s_t, a_t)),$$

where $f^{cql}$ is the regularization term

$$f^{cql} = E_s \left[ \log \sum_a \exp(Q(s_t, a_t)) - E_a(Q(s_t, a_t)) \right],$$

Conservative q-learning for offline reinforcement learning. Kumar et al. Neuirps 2020.

# Policy Regulation

- Policy regulation methods restrict the learned policy's divergence from
  - the behavior policy
  - a training dataset
  - a language model, for instance

$$V_D^\pi(s) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{s_t \sim P_t^\pi(s)} \left[ R^\pi(s_t) - \alpha D\left(\pi(\cdot|s_t), \pi_b(\cdot|s_t)\right) \right]$$

$$\min_{Q_\psi} \mathbb{E}_{\substack{(s,a,r,s') \sim \mathcal{D} \\ a' \sim \pi_\theta(\cdot|s')}} \left[ \left( r + \gamma \left( \bar{Q}(s', a') - \alpha \hat{D}\left(\pi_\theta(\cdot|s'), \pi_b(\cdot|s')\right) \right) - Q_\psi(s, a) \right)^2 \right]$$

$$\max_{\pi_\theta} \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[ \mathbb{E}_{a'' \sim \pi_\theta(\cdot|s)} \left[ Q_\psi(s, a'') \right] - \alpha \hat{D}\left(\pi_\theta(\cdot|s), \pi_b(\cdot|s)\right) \right]$$

Behavior Regularized Offline Reinforcement Learning. Wu et al. ICLR'2020.

# Outline

Part-1

- ❑ Introduction
- ❑ Interactive exercise-1
- ❑ Overview of proactive conversation agent

Part-2

- ❑ Proactive ontology expansion
- ❑ Learning to ask & topic shifting
- ❑ Counterfactual utterance generation

Part-3

- ❑ Response quality control
- ❑ **Interactive exercise-2**

# Exercise instructions

- Create pairs.
- One person acts as the user and the other acts as the agent.
- The user is given a task. **Don’t reveal it to the agent!**
- The user starts asking questions to the agent to fulfil their task.
- The agent can use **web search** or other **online means** to find the information. In addition, the agent can use any of the **techniques** we learned today while responding.
- Each response should take no more than **30 seconds**.
- Spend no more than 5 minutes on this and then reverse the roles. Repeat.
- Debriefing: 5 minutes.

# Discussion & Concluding Remarks