Unstructured Knowledge Grounding for Open-Domain Dialogues

2023.09.14

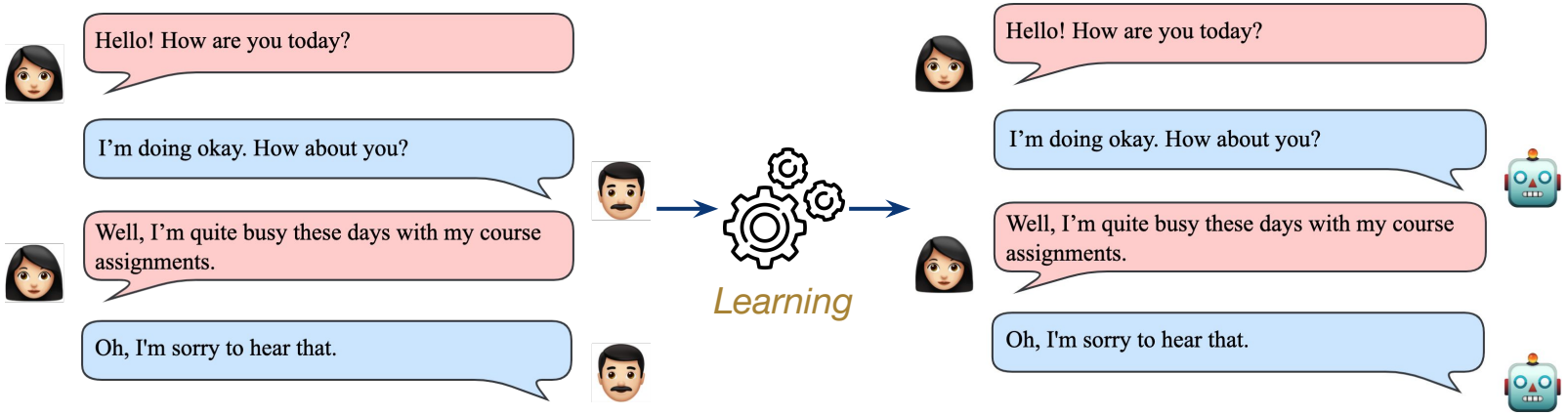
Yan Xu
Supervisor: Prof. Pascale Fung

ECE Department, HKUST

Dialogue Systems

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Dialogue systems are computer systems designed to interact with humans in natural languages.

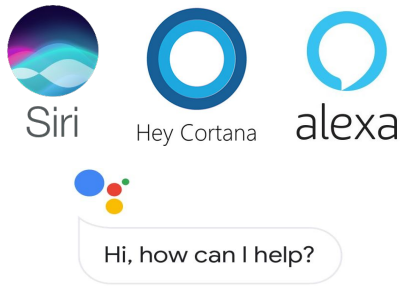


Open-Domain Dialogue Systems

Open-domain dialogue systems allows for a broader level of versatility.

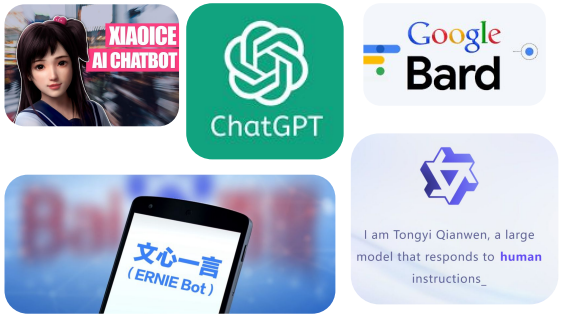
- Broader real-world applications like *social chatbots* and *metaverse-based virtual humans*.

Task-Oriented Dialogue



- *Assist users in accomplishing specific tasks with pre-defined steps*
  - Hotel reservations
  - Restaurant booking
  - Event scheduling
  - ……

Open-Domain Dialogue

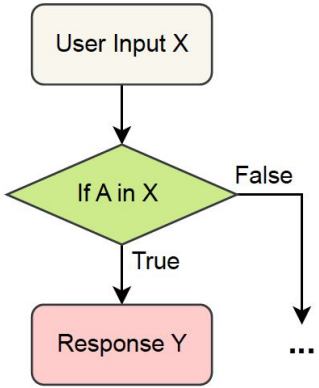


- *Broader level of versatility*
  - Everyday greetings
  - In-depth topic-specific discussions
  - ……

Open-Domain Dialogue Systems

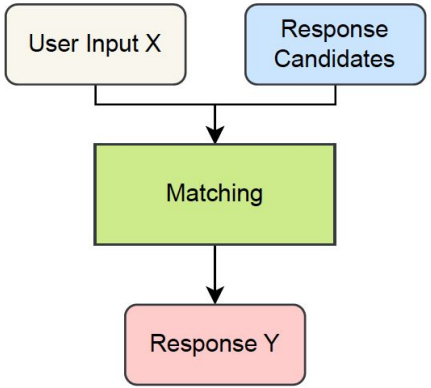
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Three categories of open-domain dialogue systems



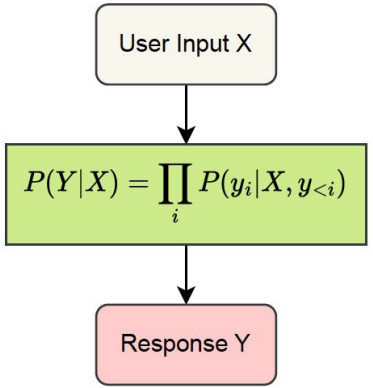
1) Rule-based

Predominantly rely on a set of predefined rules



2) Retrieval-based

Match and rank the dialogue history with candidates



3) Generation-based

Generate responses token-by-token, given the dialogue history

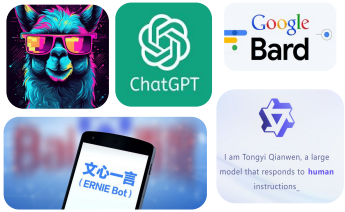
- *Great flexibility and scalability in building intelligent chatbots*
- *Minimal reliance on hand-crafting*

Generative Open-Domain Dialog Systems

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CAiRE Chatbot (CAiRE HKUST, 2019)
A neural empathetic chatbot based on PLM

Meena: Towards a Human-like Open-Domain Chatbot (Google, 2020)
2.6B parameters; 341GB data.
BlenderBot (Meta, 2020)
9.4B parameters; 1.5B conversations.



ChatGPT (OpenAI, 2022)
Pre-training + RLHF
GPT-4 (OpenAI, 2023)
> 175B parameters
Bard (Google, 2023)
LLaMA2 (Meta, 2023)
70B parameters
ERNIE Bot (Baidu, 2023)

DialoGPT (Microsoft, 2019) the first large-scale pre-trained dialogue model based on GPT-2.
762M parameters; 147M conversations.

LaMDA (Google, 2022)
137B parameters; 1.12B dialogue + 2.97B documents.
BlenderBot 3.0 (Meta, 2022)
175B parameters;

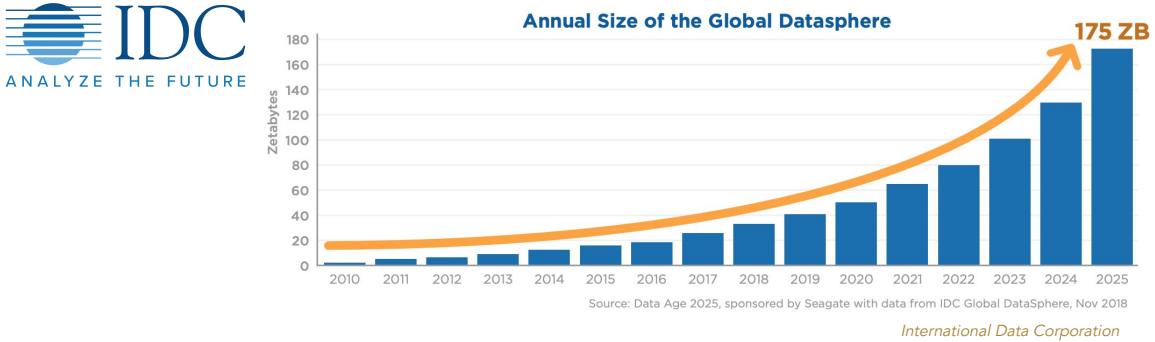
- *Scaling up* has shown improvement in dialogue generation.
  - Language capabilities are improved with implicit knowledge inside their parameters
  - Natural, fluent, and plausible responses
- ChatGPT and GPT-4 demonstrate a superior ability for interaction, solely relying on the *parametric knowledge from LLMs*.
- The promise of large language model (LLM)-based chatbots has been proven.

Limitations of LLM-based Dialogue Models

Despite the strong capabilities of LLM-based dialogue models, challenges still remain: *They are not fully reliable.*

1. Lack of New Knowledge

- Not exposed to *new knowledge*: World knowledge is *increasing* and frequently updated. E.g. GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its pre-training data cuts off in September 2021.



Limitations of LLM-based Dialogue Models

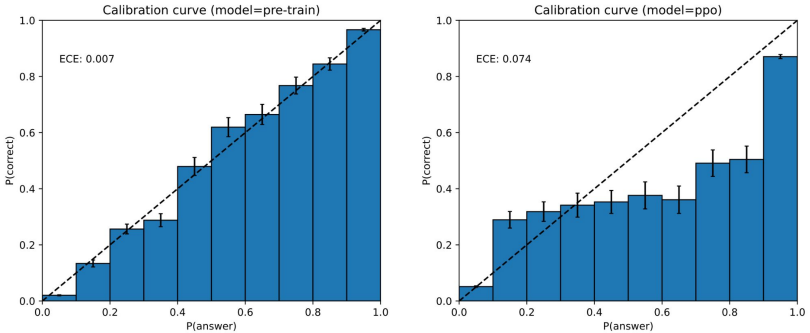
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Despite the strong capabilities of LLM-based dialogue models, challenges still remain: *They are not fully reliable.*

1. Lack of New Knowledge

- Not exposed to *new knowledge*: World knowledge is *increasing* and frequently updated. E.g. GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its pre-training data cuts off in September 2021.
- *Catastrophic forgetting* of the knowledge obtained from pre-training through post-training.

before RLHF



after RLHF

"Interestingly, the pre-trained model is highly calibrated (its predicted confidence in an answer generally matches the probability of being correct). However, after the post-training process, the calibration is reduced."[1]

[1] "GPT-4 Technical Report." OpenAI, 2023.

RLHF: Reinforcement Learning with Human Feedback

Limitations of LLM-based Dialogue Models

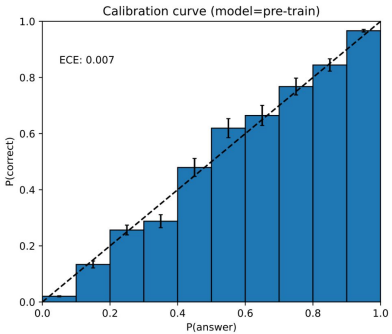
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

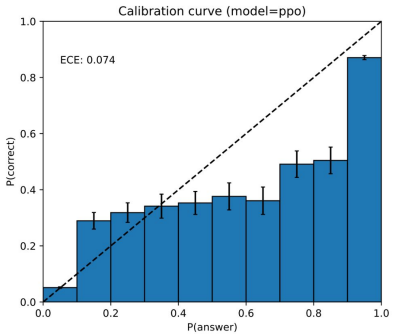
Despite the strong capabilities of LLM-based dialogue models, challenges still remain: *They are not fully reliable.*

1. Lack of New Knowledge

- Not exposed to *new knowledge*: World knowledge is *increasing* and frequently updated. E.g. GPT-4 generally lacks knowledge of events that have occurred after the vast majority of its pre-training data cuts off in September 2021.
- *Catastrophic forgetting* of the knowledge obtained from pre-training through post-training.
- It is not feasible to frequently update the model parameters to *adapt to new emergent knowledge*.

before RLHF





after RLHF

[1] "GPT-4 Technical Report." OpenAI, 2023.

RLHF: Reinforcement Learning with Human Feedback

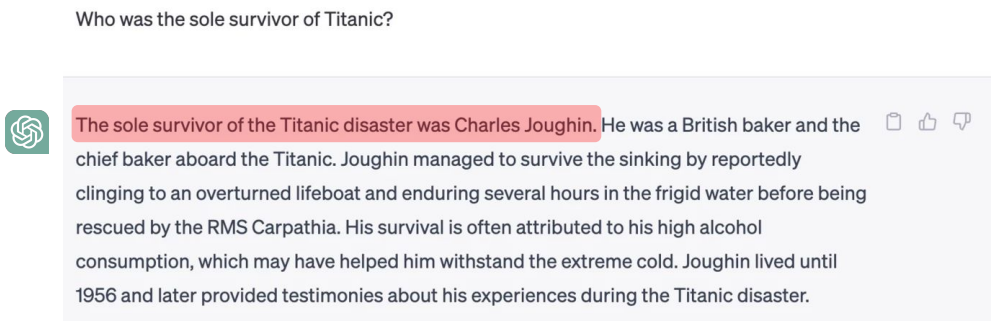
Limitations of LLM-based Dialogue Models

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Despite the strong capabilities of LLM-based dialogue models, challenges still remain: *They are not fully reliable.*

2. Hallucination

- GPT-4 has the tendency to “hallucinate”, i.e. “produce content that is plausible-sounding but nonsensical or untruthful in relation to certain sources.” [1]
- It generates more extrinsic hallucinations from its parametric memory as it does not have access to external knowledge sources.[2]



Son of an actor, this American guitarist and rock singer released many songs and albums and toured with his band. His name is "Elvis" what?

Perkins
Presley ← choice
Elvis Presley
His name is Elvis Presley

[1] “GPT-4 Technical Report.” OpenAI, 2023.
[2] “A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.” Bang, Yejin, et al., AACL 2023.

Limitations of LLM-based Dialogue Models

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Despite the strong capabilities of LLM-based dialogue models, challenges still remain: *They are not fully reliable.*

- Lack of New Knowledge
- Hallucination

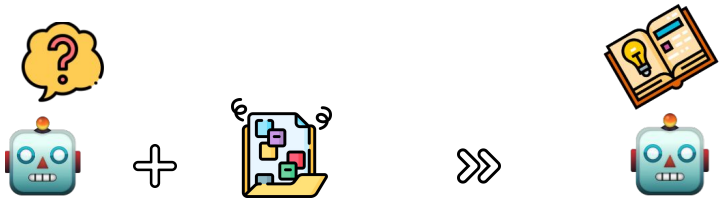
*Open-domain dialogue systems still need to further ground on **knowledge from external sources***

[1] "GPT-4 Technical Report." OpenAI, 2023.
[2] "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity." Bang, Yejin, et al., AACL 2023.

Our Goal

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

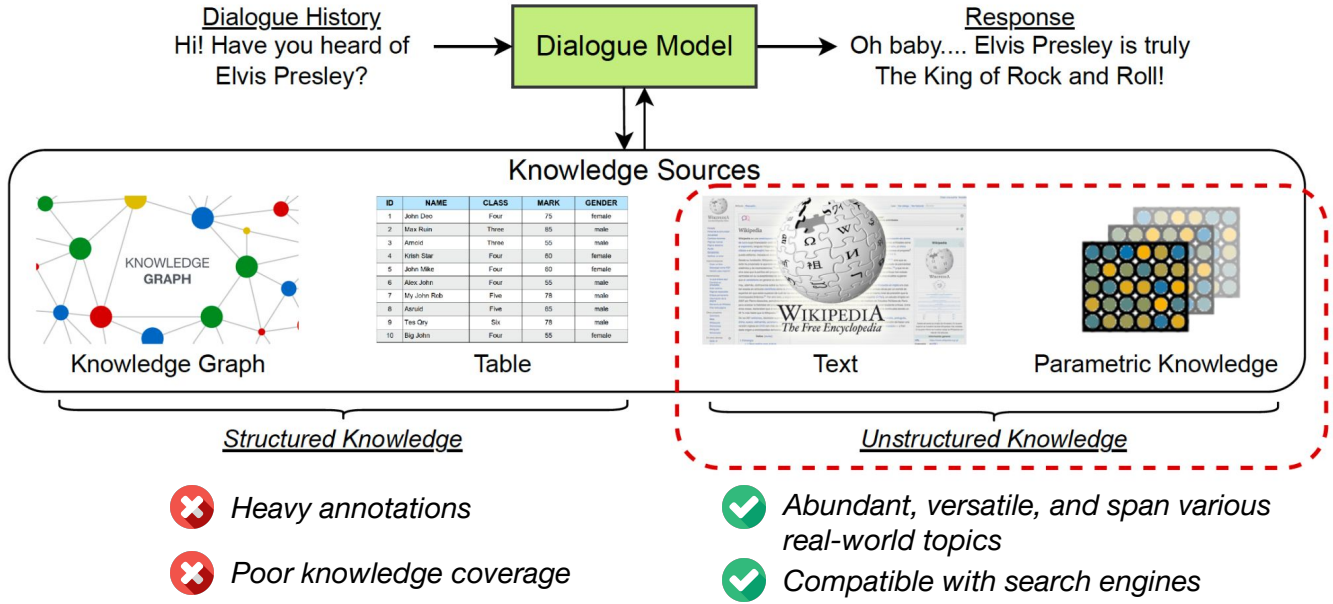
Goal: Building knowledgeable and trustworthy open-domain dialogue models with unstructured knowledge grounding



Our Goal

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Goal: Building knowledgeable and trustworthy open-domain dialogue models with unstructured knowledge grounding



Research Questions

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Goal: Building knowledgeable and trustworthy open-domain dialogue models with unstructured knowledge grounding

1 How to build an adaptive open-domain dialogue system with unstructured knowledge grounding?

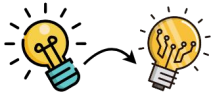
Research Questions

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Goal: Building knowledgeable and trustworthy open-domain dialogue models with unstructured knowledge grounding

1. How to build an adaptive open-domain dialogue system with unstructured knowledge grounding?

2. How to acquire more transferable skills for unstructured knowledge grounding?



Research Questions

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Goal: Building knowledgeable and trustworthy open-domain dialogue models with unstructured knowledge grounding

1. How to build an adaptive open-domain dialogue system with unstructured knowledge grounding?

2. How to acquire more transferable skills for unstructured knowledge grounding?

3. How to generate responses that are more faithful to external knowledge?

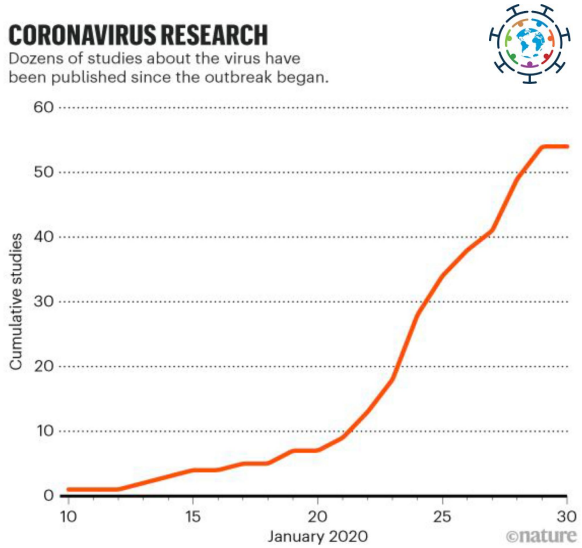
Knowledge Grounding for Adaptive Open-Domain Dialogue Systems

"Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters", Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, Pascale Fung, ACL2022 DialDoc

Adaptiveness: Motivations

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- GPT-4 and other LLM-based dialogue models *lack new knowledge*, notwithstanding the fact that world knowledge is *increasing* over time.
- ***New topics*** emerge with new knowledge, e.g. COVID-19.
- LLM-based dialogue models are *unreliable* in handling new topics because they are *not exposed to such knowledge* during pre-training.
- Further explorations are required to tackle the challenge of ***adapting*** open-domain dialogue models to ***new topics***. However, ***no prior work*** has been conducted.



Our Path to Solution

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

We investigate

- How to build an open-domain dialogue system so that it is more knowledgeable about new topics.
- How to adapt a GPT-2[17] model for open-domain dialogues to handle conversations in Wikipedia topics from Wizard of Wikipedia and CMU DoG datasets.

A dialogue model that generates more informative responses in certain topics is more knowledgeable about such topics without external knowledge as inputs

[17] "Language Models are Unsupervised Multitask Learners." Radford, Alec et al., OpenAI blog, 2019.

Our Path to Solution

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

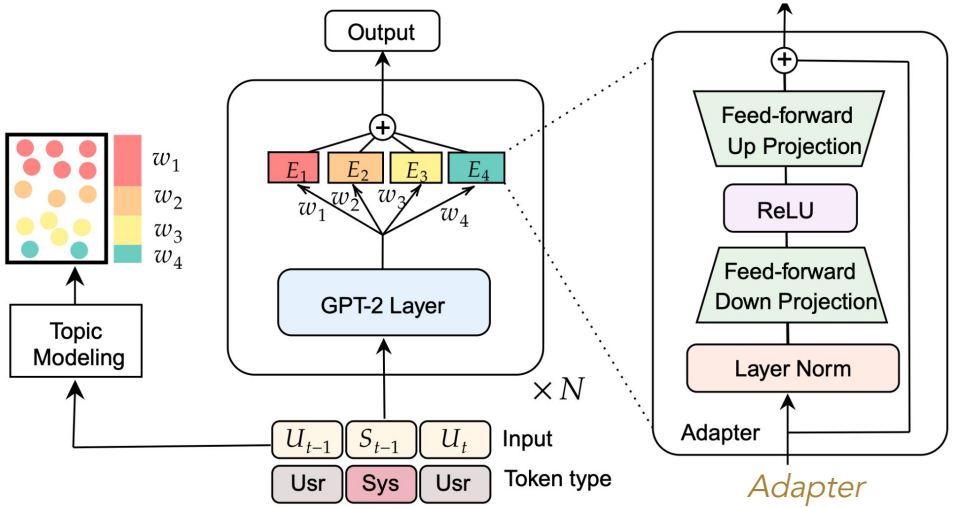
We investigate

- How to build an open-domain dialogue system so that it is more knowledgeable about new topics.
- How to adapt a GPT-2[17] model for open-domain dialogues to handle conversations in Wikipedia topics from Wizard of Wikipedia and CMU DoG datasets.

We propose

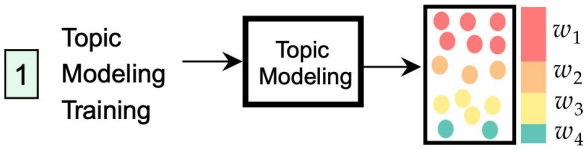
1. Inject Wikipedia knowledge into parameters of GPT-2 model with *lightweight adapters* for certain topics;
2. Conversational style pre-training and a contextualized topic model *amplify the utility* of injected knowledge.

KnowExpert: Model



- ***Knowledge Injection***: Pre-train lightweight adapters and insert adapters upon GPT2 layers, where the adapters act as *knowledge experts* for certain topics.
- ***Response Generation***: Generate a response with the mixture of knowledge experts whose weights are determined based on the relevance to the topic.

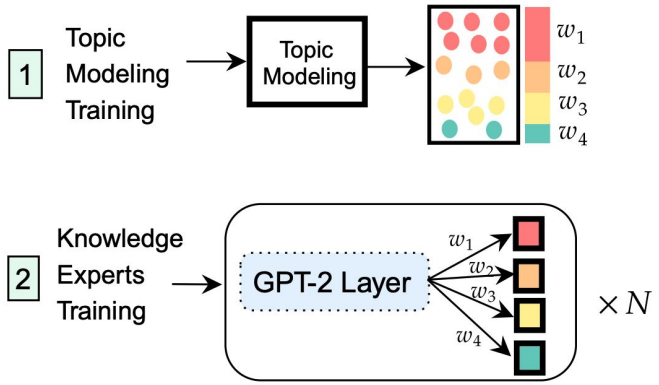
KnowExpert: Learning Procedure



Our training follows a three-step approach:

1. Train the ***contextualized topic model*** using knowledge corpus and cluster the knowledge sentences based on the topics.

KnowExpert: Learning Procedure



Our training follows a three-step approach:

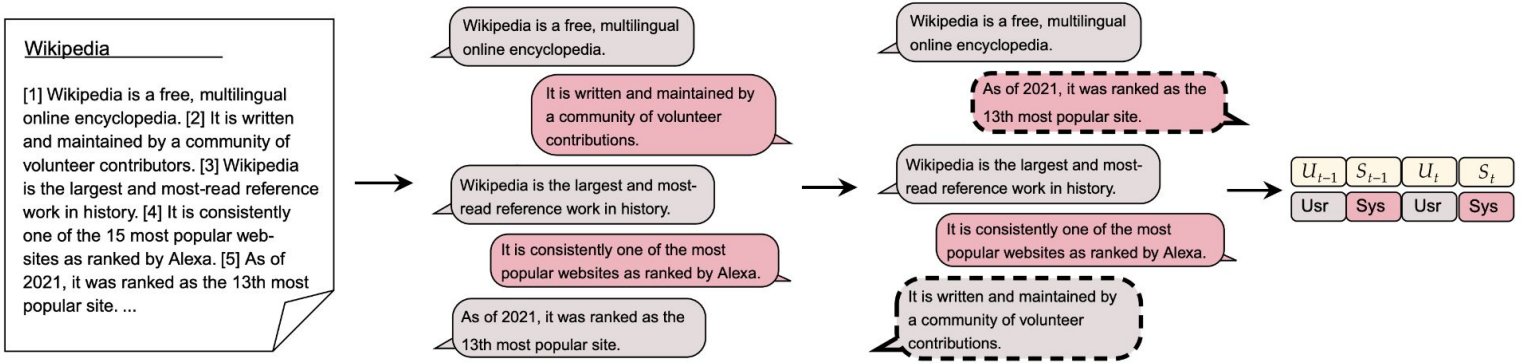
1. Train the ***contextualized topic model*** using knowledge corpus and cluster the knowledge sentences based on the topics.
2. Train different ***topic-specific knowledge experts*** with the frozen GPT-2 backbone with knowledge sentences in different clusters in a *pseudo-conversational style*.

$$\mathcal{L}_{\mathcal{K}^l} = -\frac{1}{|\mathcal{K}^l|} \sum_{K \in \mathcal{K}^l} \log p_{\theta_{E_l}}(K)$$

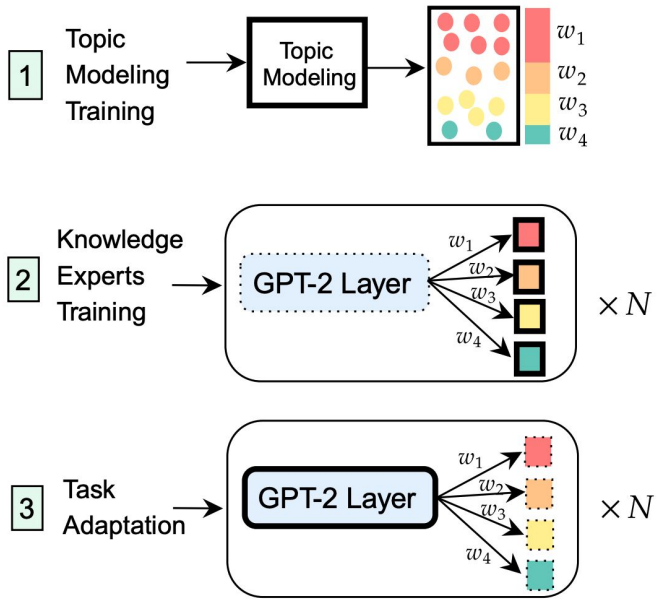
KnowExpert: Learning Procedure

The procedure of converting an article in the knowledge corpus (e.g., a Wikipedia article) into the pseudo-conversation style.

1. Sentence tokenization
2. Random permutation
3. Dialogue-oriented pre-training



KnowExpert: Learning Procedure



3. Finetune the ***backbone GPT2 model*** with the frozen knowledge experts using target dialogue datasets.

$$\mathcal{L}_{\text{Task}} = -\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{w}\sim q_\phi}\log p_{\theta_K}(R^n|H_t^n,\boldsymbol{w}^n)$$

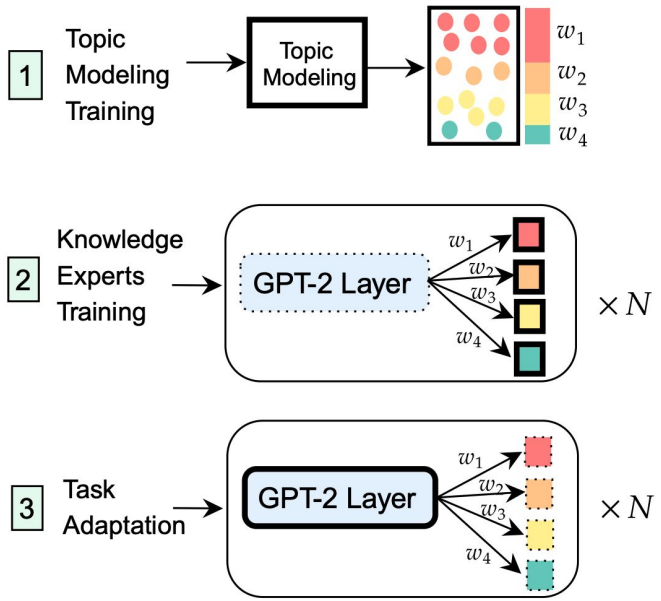
   a. *How do we **decide the weights over different knowledge experts** for each dialogue sample?*

   Topic model assesses the topic relevance

      i. **Observation 1:** There is a discrepancy between dialogue data and Wikipedia sentences.

      ii. **Observation 2:** It is more accurate to predict the cluster distribution with both dialogue history and the response as inputs.

KnowExpert: Learning Procedure



3. Finetune the ***backbone GPT2 model*** with the frozen knowledge experts using target dialogue datasets.

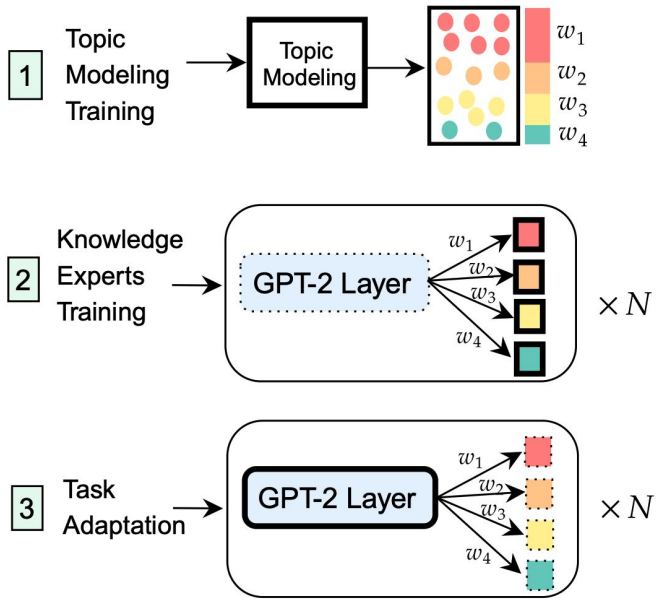
$$\mathcal{L}_{\text{Task}} = -\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{w}\sim q_\phi}\log p_{\theta_K}(R^n|H_t^n, \boldsymbol{w}^n)$$

a. *How do we **decide the weights over different knowledge experts** for each dialogue sample?*

iii. Further finetune the topic model to minimize the **MSE loss** between the output distribution when using `dialogue history+response` and `dialogue history only` as inputs.

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(q_\phi(H^i, R^i) - \pi_\theta(H^i))^2$$

KnowExpert: Learning Procedure



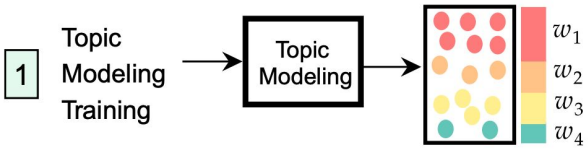
3. Finetune the ***backbone GPT2 model*** with the frozen knowledge experts using target dialogue datasets.

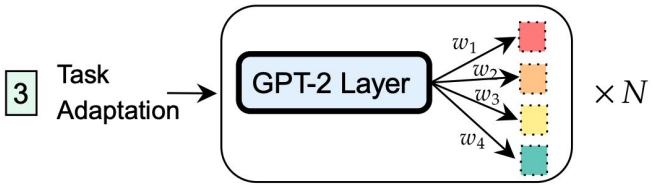
$$\mathcal{L}_{\text{Task}} = -\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\boldsymbol{w}\sim q_\phi}\log p_{\theta_K}(R^n|H_t^n, \boldsymbol{w}^n)$$

b. *How do we **leverage the weights over different knowledge experts** during task adaptation?*
   i. weighted-sum over all the knowledge experts (KnowExpert$_w$)
   ii. only use the knowledge expert with the highest weight (KnowExpert$_o$)

$$\boldsymbol{w} = \begin{cases} q_\phi(H, R), & \text{if weighted-sum,} \\ \mathbb{1}(q_\phi(H, R)), & \text{if one-hot.} \end{cases}$$

KnowExpert: Generation Process





During the generation process, given one data sample:

1. Leverage the finetuned topic model to predict the topic cluster distributions;

$$\boldsymbol{w} = \pi_\theta(H)$$

2. Conduct dialogue response generation with the KnowExpert model under either a weighted-sum or a one-hot setting.

$$\hat{R} \sim \begin{cases} p_{\theta_K}(R|H, \boldsymbol{w}), & \text{if weighted-sum,} \\ p_{\theta_K}(R|H, \mathbb{1}(\boldsymbol{w})), & \text{if one-hot.} \end{cases}$$

Experiments

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Datasets

- Wizard of Wikipedia (WoW) [3] is a KGD dataset across 1,365 topics
  - 18,430/1,948/965/968 train/valid/test seen/test unseen dialogues
  - Ground on Wikipedia articles
- CMU DoG [4] is a KGD dataset in the movie domain
  - 3,373/229/619 train/valid/test dialogues
  - Ground on Wikipedia articles

Baselines

- Retrieval-free baselines:
  - GPT-2f
    - dialogue history -> response
  - KE-Blender [5]
    - knowledge-enhanced fine-tuning
- Retrieval-based baselines:
  - DRD [6]
  - ZRGKC [7]
  - KnowledGPT [8]
    - one of the SOTA models

[3] "Wizard of Wikipedia: Knowledge-Powered Conversational Agents". Dinan, Emily, et al. ICLR 2018.
[4] "A Dataset for Document Grounded Conversations". Zhou, K., Prabhumoye, S., & Black, A. W. EMNLP 2018.
[5] "Knowledge Enhanced Fine-Tuning for Better Handling Unseen Entities in Dialogue Generation". Cui, Leyang et al. EMNLP 2021.
[6] "Low-Resource Knowledge-Grounded Dialogue Generation". Zhao, Xueliang, et al. ICLR 2020.
[7] "Zero-Resource Knowledge-Grounded Dialogue Generation". Li, Linxiao et al. NeurIPS 2020.
[8] "Knowledge-Grounded Dialogue Generation with Pre-trained Language Models". Zhao, Xueliang et al. EMNLP 2020.

Results: Automatic Evaluation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- KnowExpert performs comparably with the retrieval-based approaches, especially on seen domain.

Model		WoW Seen				WoW Unseen				CMU_DoG	
		PPL↓	F1↑	Dist-1↑	Dist-2↑	PPL↓	F1↑	Dist-1↑	Dist-2↑	PPL↓	F1↑
Retrieval-based Approach	DRD	23.0	18.0	-	-	25.6	16.5	-	-	54.4	10.7
	ZRGKG	40.4	18.7	5.4	22.5	41.5	18.6	3.4	15.6	53.5	12.5
	GPT-2_{trunc}	14.6	18.7	-	-	16.9	18.3	-	-	18.6	10.8
	KnowledGPT	19.2	**22.0**	8.9	36.2	22.3	**20.5**	6.0	23.8	20.6	**13.5**
Retrieval-free Approach	GPT-2_f	18.8	17.0	4.9	21.1	21.0	16.3	3.9	16.8	17.8	11.8
	KE-Blender†	15.5	17.0	-	-	18.4	**16.7**	-	-	-	-
	KnowExpert$_w$+causal	15.2	18.4	6.4	26.4	20.0	16.6	4.9	20.4	16.8	12.1
	KnowExpert$_o$ (ours)	16.0	18.4	6.6	27.2	21.2	16.6	**5.2**	**21.6**	17.8	12.1
	KnowExpert$_w$ (ours)	15.3	**18.7**	**6.8**	**27.9**	20.1	**16.7**	**5.2**	21.2	17.2	**12.5**

PPL: Exponent of negative likelihood of the target response;
F1: Unigram overlap between the generated and target responses;
Dist-1/2: Diversity measure by the portion of distinct uni/bi-grams in the generated responses.

Results: Automatic Evaluation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- KnowExpert shows an advantage over retrieval-free baselines

Model		WoW Seen				WoW Unseen				CMU_DoG	
		PPL↓	F1↑	Dist-1↑	Dist-2↑	PPL↓	F1↑	Dist-1↑	Dist-2↑	PPL↓	F1↑
Retrieval-based Approach	DRD	23.0	18.0	-	-	25.6	16.5	-	-	54.4	10.7
	ZRGKG	40.4	18.7	5.4	22.5	41.5	18.6	3.4	15.6	53.5	12.5
	GPT-2_{trunc}	14.6	18.7	-	-	16.9	18.3	-	-	18.6	10.8
	KnowledGPT	19.2	**22.0**	8.9	36.2	22.3	**20.5**	6.0	23.8	20.6	**13.5**
Retrieval-free Approach	GPT-2_{f}	18.8	17.0	4.9	21.1	21.0	16.3	3.9	16.8	17.8	11.8
	KE-Blender†	15.5	17.0	-	-	18.4	**16.7**	-	-	-	-
	KnowExpert$_{\text{w}}$+causal	15.2	18.4	6.4	26.4	20.0	16.6	4.9	20.4	16.8	12.1
	KnowExpert$_{\text{o}}$ (ours)	16.0	18.4	6.6	27.2	21.2	16.6	**5.2**	**21.6**	17.8	12.1
	KnowExpert$_{\text{w}}$ (ours)	15.3	**18.7**	**6.8**	**27.9**	20.1	**16.7**	**5.2**	21.2	17.2	**12.5**

PPL: Exponent of negative likelihood of the target response;
F1: Unigram overlap between the generated and target responses;
Dist-1/2: Diversity measure by the portion of distinct uni/bi-grams in the generated responses.

Results: Automatic Evaluation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- KnowExpert$_w$ shows consistently better performance over KnowExpert$_o$.
- Without converting the knowledge corpus into pseudo-conversation style, the performance of the model (KnowExpert$_w$+causal) drops even below that of KnowExpert with the one-hot setting, which shows the importance of the pseudo-conversation style pre-training.

Model		WoW Seen				WoW Unseen				CMU_DoG	
		PPL↓	F1↑	Dist-1↑	Dist-2↑	PPL↓	F1↑	Dist-1↑	Dist-2↑	PPL↓	F1↑
Retrieval-based Approach	DRD	23.0	18.0	-	-	25.6	16.5	-	-	54.4	10.7
	ZRGKG	40.4	18.7	5.4	22.5	41.5	18.6	3.4	15.6	53.5	12.5
	GPT-2$_{\text{trunc}}$	14.6	18.7	-	-	16.9	18.3	-	-	18.6	10.8
	KnowledGPT	19.2	**22.0**	8.9	36.2	22.3	**20.5**	6.0	23.8	20.6	**13.5**
Retrieval-free Approach	GPT-2$_f$	18.8	17.0	4.9	21.1	21.0	16.3	3.9	16.8	17.8	11.8
	KE-Blender†	15.5	17.0	-	-	18.4	**16.7**	-	-	-	-
	KnowExpert$_w$+causal	15.2	18.4	6.4	26.4	20.0	16.6	4.9	20.4	16.8	12.1
	KnowExpert$_o$ (ours)	16.0	18.4	6.6	27.2	21.2	16.6	**5.2**	**21.6**	17.8	12.1
	KnowExpert$_w$ (ours)	15.3	**18.7**	**6.8**	**27.9**	20.1	**16.7**	**5.2**	21.2	17.2	**12.5**

PPL: Exponent of negative likelihood of the target response;
F1: Unigram overlap between the generated and target responses;
Dist-1/2: Diversity measure by the portion of distinct uni/bi-grams in the generated responses.

Results: Human Evaluation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Winning Rate (%)	WoW Seen		WoW Unseen	
Models	Info.	Human.	Info.	Human.
KnowExpert$_w$ vs. GPT-2$_f$	**57.68**	48.69	**59.26**	**56.13**
KnowExpert$_o$ vs. GPT-2$_f$	**64.46**	**54.42**	**55.88**	**53.67**

KnowExpert succeeds to generate more informative responses without losing human-likeness, compared with the GPT-2f baseline.

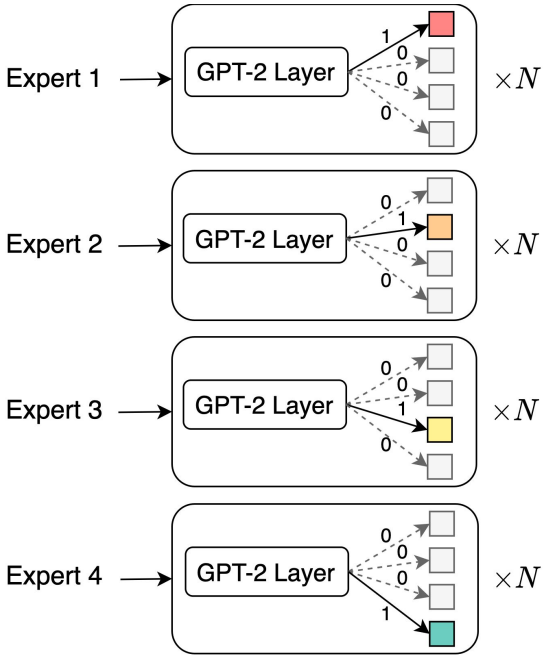
Informativeness: How informative the generated responses are, based on the amount of new information introduced into the conversations and the factualness of the responses;
Human-Likeness: Fluency and coherence of the generated responses;
Results in bold: Statistically significant based on a pair-wise individual t-test ($p<0.05$).

Analysis: Case Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

How do different knowledge experts affect the final generation?

Context		User: Orc. System: Orcs are cool fictional humanoid beings. User: Yeah, I've seen them in a lot of things like magic and dnd.	
Generated responses with single knowledge expert in KnowExpert$_{w}$ ($L = 4$)	Expert 1	Do you know about the orcs? They are native to the Italian peninsula. ***Topics of Cluster 1**: east, river, south, state, city, area, island, ...*	✗
	Expert 2	They are a subgenre of "art games" that are a subgenre of video games. ***Topics of Cluster 2**: rock, band, music, album, football, single, ...*	✗
	Expert 3	Orcs are cool, they are a subspecies of elves in the warcraft universe. ***Topics of Cluster 3**: fiction, story, characters, novel, film, stars, ...*	✓
	Expert 4	They are a legendary race that are native to the americas. ***Topics of Cluster 4**: bon, bucks, rutgers, canberra, ivy, nets, ...*	✗
KnowExpert$_{w}$		They are a fictional humanoid creature from the "dungeons & dragons" fantasy roleplaying game.	



Analysis: Case Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

How do different knowledge experts affect the final generation?

Context		User: Orc. System: Orcs are cool fictional humanoid beings. User: Yeah, I've seen them in a lot of things like magic and dnd.	
Generated responses with single knowledge expert in KnowExpert$_w$ ($L = 4$)	Expert 1	Do you know about the orcs? They are native to the Italian peninsula. ***Topics of Cluster 1**: east, river, south, state, city, area, island, ...*	✗
	Expert 2	They are a subgenre of "art games" that are a subgenre of video games. ***Topics of Cluster 2**: rock, band, music, album, football, single, ...*	✗
	Expert 3	Orcs are cool, they are a subspecies of elves in the warcraft universe. ***Topics of Cluster 3**: fiction, story, characters, novel, film, stars, ...*	✓
	Expert 4	They are a legendary race that are native to the americas. ***Topics of Cluster 4**: bon, bucks, rutgers, canberra, ivy, nets, ...*	✗
KnowExpert$_w$		They are a fictional humanoid creature from the "dungeons & dragons" fantasy roleplaying game.	

Our knowledge experts tend to focus on the *topics* to which the *knowledge documents they are trained on belong to*.

A knowledge expert whose topics are more similar to the topic of the dialogue tends to generate *more factual* responses.

Tolkien's concept of orcs has been adapted into the fantasy fiction of other authors, and into games of many different genres such as Dungeons & Dragons, Magic: The Gathering, and Warcraft.

Analysis: Case Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

How do different knowledge experts affect the final generation?

Context		User: Orc. System: Orcs are cool fictional humanoid beings. User: Yeah, I've seen them in a lot of things like magic and dnd.	
Generated responses with single knowledge expert in KnowExpert$_w$ ($L = 4$)	Expert 1	Do you know about the orcs? They are native to the Italian peninsula. ***Topics of Cluster 1**: east, river, south, state, city, area, island, ...*	✗
	Expert 2	They are a subgenre of "art games" that are a subgenre of video games. ***Topics of Cluster 2**: rock, band, music, album, football, single, ...*	✗
	Expert 3	Orcs are cool, they are a subspecies of elves in the warcraft universe. ***Topics of Cluster 3**: fiction, story, characters, novel, film, stars, ...*	✓
	Expert 4	They are a legendary race that are native to the americas. ***Topics of Cluster 4**: bon, bucks, rutgers, canberra, ivy, nets, ...*	✗
KnowExpert$_w$		They are a fictional humanoid creature from the "dungeons & dragons" fantasy roleplaying game.	

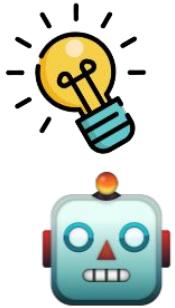
The *mixture-of-experts* approach ensures a better model performance. The generated response of KnowExpert$_w$ is more *on-topic and accurate* thanks to leveraging the weighted sum of the experts.

Summary: Adaptiveness

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Unstructured knowledge grounding for **adaptive** open-domain dialogue systems

- Adaptive to *new topics* which improves the *reliability*
- The *first* attempt to inject unstructured knowledge with lightweight adapters for KGD tasks
- *More knowledgeable* by generating more informative responses without an explicit retrieval step
- *Orthogonal* with retrieval-augmented methods



Knowledge Grounding for Transferable Open-Domain Dialogue Systems

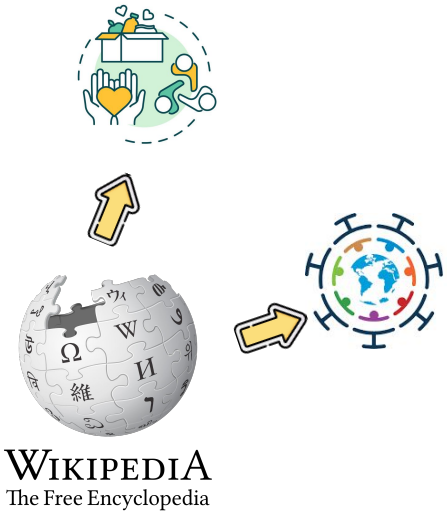
"CAiRE in DialDoc21: Data Augmentation for Information-Seeking Dialogue System", Yan Xu, Etsuko Ishii, , Genta Indra Winata, Zhaojiang Lin, Andrea Madotto, Zihan Liu, Peng Xu, Pascale Fung, ACL2021 DialDoc

Transferability: Motivations

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- ***Limited high-quality KGD data*** makes it hard to build robust knowledge-grounded dialogue (KGD) systems.
- KGD systems that are designed for *general purpose* and grounded on *Wikipedia* falter in niche, low-resource domains.

- There is a *flourishing demand* for chatbots to handle information-seeking queries in niche domains like *social welfare and COVID-19 policies*.
- Building such KGD systems heavily relies on the ***transferability*** of the *knowledge grounding skills* from general domains to other domains, which is undeniably *critical* yet remains *underexplored*.



Our Path to Solution

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

We investigate

- How to obtain *transferable* knowledge grounding skills with a keen emphasis on information-seeking dialogues in niche, low-resource domains.

We propose QA4KGD

- We employ *modularity* and *decompose* the open-domain KGD system into two sub-modules.
- Our strategy highlights the power of *task-wise pre-training* coupled with *multi-task learning (MTL)*. They augment the models to learn more transferable task skills with less sensitivity to the domain differences.

QA4KGD: Question Answering for Knowledge Grounded Dialouge

QA4KGD: Task Decomposition

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

To Learn a model f_θ to generate an informative response R to fulfill users' needs for the domain-specific information grounded with natural language documents.

$$f_\theta(H, K) \rightarrow R$$

natural language document

QA4KGD: Task Decomposition

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

To Learn a model f_θ to generate an informative response R to fulfill users' needs for the domain-specific information grounded with natural language documents.

$$f_\theta(H, K) \rightarrow R$$

We propose to decompose the complex task into two modules: $\theta = \{\alpha, \beta\}$

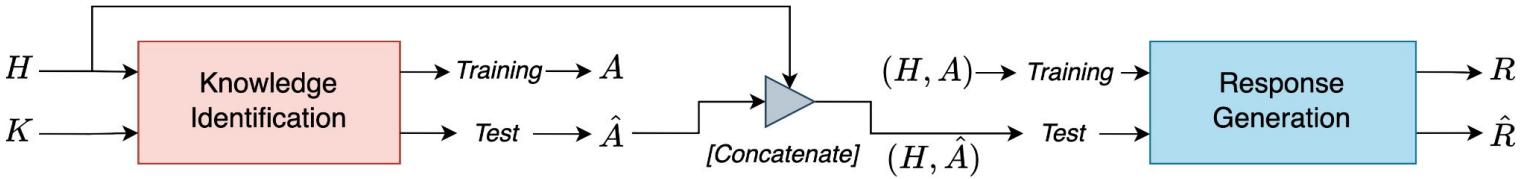
Knowledge Identification

$$f_\alpha(H, K) \rightarrow A$$

Response Generation

$$f_\beta(H, A) \rightarrow R$$

natural language document



An illustration of the QA4KGD framework

QA4KGD: Task Decomposition

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

To Learn a model f_θ to generate an informative response R to fulfill users' needs for the domain-specific information grounded with natural language documents.

$$f_\theta(H, K) \rightarrow R$$

We propose to decompose the complex task into two modules: $\theta = \{\alpha, \beta\}$

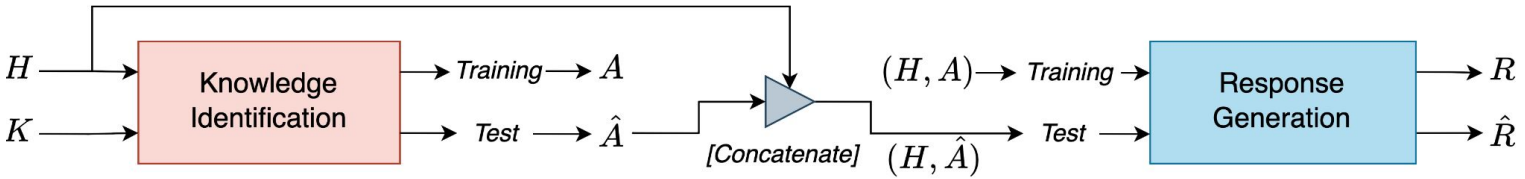
Knowledge Identification

$$f_\alpha(H, K) \rightarrow A$$

Similar to QA formulation

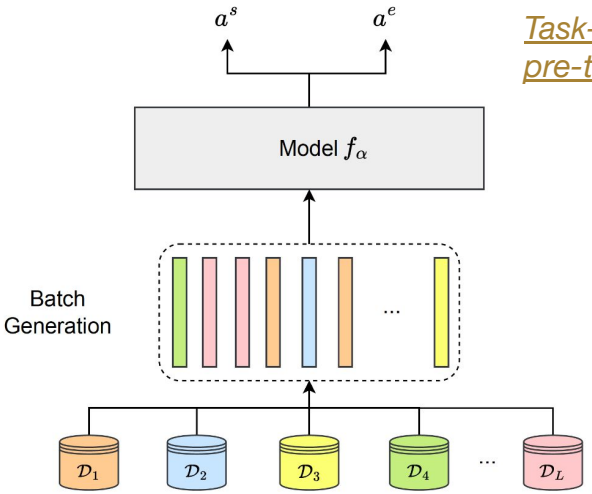
Response Generation

$$f_\beta(H, A) \rightarrow R$$



An illustration of the QA4KGD framework

QA4KGD: Knowledge Identification



(a) Multi-task learning method of QA model

Learning Method

Task-wise pre-training

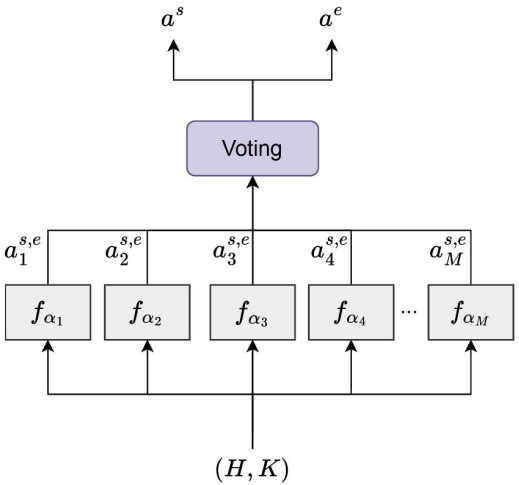
1. Pre-training on general QA tasks with multi-task learning
2. Continual learning on CQA datasets with pre-trained QA models
3. Fine-tuning on the target domain

$$\mathcal{L}(\alpha) = -\frac{1}{N}\sum_{n=1}^{N} \log p_\alpha(a^{s_n}|H^n, K^n) + \log p_\alpha(a^{e_n}|H^n, K^n)$$

Answer Prediction

$$\hat{A} = \arg\max_{(a^s, a^e)} p_\alpha(a^s) p_\alpha(a^e), \text{ where } \quad a^s \leq a^e$$

QA4KGD: Knowledge Identification



(b) Ensemble scheme

Ensemble Scheme

- One prediction (a^s, a^e) as a unit
- The most frequent span as the final prediction

Post-Processing

- Take the complete sentences for response generation

QA4KGD: Response Generation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Learning with oracle knowledge

- Task-wise pre-training with the general KGD datasets

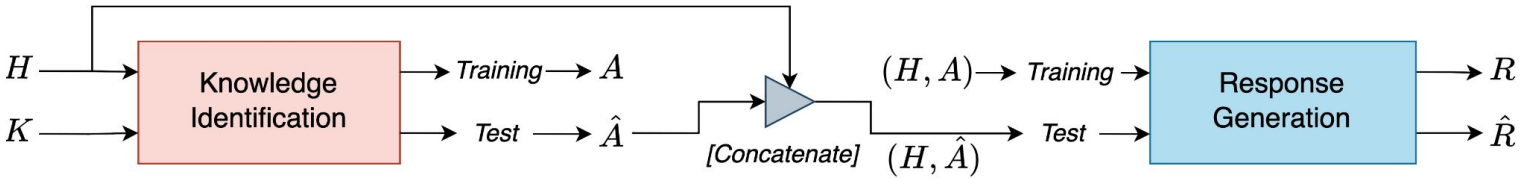
$$\mathcal{L}_{\mathrm{PT}}(\beta) = -\frac{1}{N}\sum_{n=1}^{N}\log p_\beta(R^n|H^n, K^n)$$

- Fine-tuning on the target domain

$$\mathcal{L}_{\mathrm{FT}}(\beta) = -\frac{1}{N}\sum_{n=1}^{N}\log p_\beta(R^n|H^n, A^n)$$

Generation with predicted knowledge

$$\hat{R} = f_\beta(H, \hat{A})$$



Experiments

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Datasets

- *Target Domain*
  - Doc2Dial [9]
    - 3,474/661/198/787 train/valid/testdev/test dialogues
    - Testdev set: *social welfare*; Test set contains social welfare and *unseen* topic (COVID-19).
- *Task-wise pre-training*
  - General QA datasets
    - MRQA [10] is a collection of six general QA datasets, including SQuAD, NewsQA, TriviaQA, SearchQA, HotpotQA, and NaturalQuestions.
  - CQA datasets
    - CoQA [11], QuAC [12], DoQA [13]
  - KGD dataset
    - Wizard of Wikipedia (WoW) [3]
      - 18,430 dialogues in the training set

[3] "Wizard of Wikipedia: Knowledge-Powered Conversational Agents". Dinan, Emily, et al. ICLR 2018.
[8] "doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset". Feng, Song, et al. EMNLP 2020.
[10] "MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension". Fisch, Adam, et al. EMNLP 2019 MRQA.
[11 "CoQA: A Conversational Question Answering Challenge". Reddy, Siva, et al. TACL 2019.
[12] "QuAC: Question Answering in Context". Choi, Eunsol, et al. EMNLP 2018.
[13] "DoQA - Accessing Domain-Specific FAQs via Conversational QA". Campos, Jon Ander. ACL 2020.

Results: Automatic Evaluation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Knowledge Identification

Model	# of ckpt	EM	F1
Baselines			
BERT-QA†	-	45.45	59.51
RoBERTa$_{large}$	-	58.08	72.17
Ours QA4KGD$_{KI}$	-	**65.66**	**78.23**

Results on the testdev set

Model	EM	F1
Baselines		
BERT-QA†	35.80	52.60
RoBERTa$_{large}$	53.24	68.40
Ours QA4KGD$_{KI}$	**64.42**	**77.27**

Results on the test set

Response Generation

Model	SacreBLEU		
	Valid	Testdev	Test
Baselines			
BART$_{large}$†	-	-	17.60
KnowledGPT‡	4.79	-	-
QA4KGD (Oracle)	47.74	-	-
- WoW pre-training	19.63	-	-
- Doc2Dial (Zero-shot)	27.95	-	-
Ours QA4KGQ	-	-	**39.88**
- Post-proc.	**39.72**	**36.61**	38.72(↓1.16)

- QA4KGD outperforms the baseline models by a large margin on both testdev and test sets.

EM:The portion of predictions exactly the same as gold answers;
F1: Unigram overlap between the predicted and target answers;
SacreBLEU: Unigram overlap between the generated and target responses with length penalty.

Analysis: Ablation Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Knowledge Identification

Model	# of ckpt	EM	F1
$QA4KGD_{KI}$	-	**65.66**	**78.23**
- Post-proc.	-	65.15(↓0.51)	78.46(↑0.23)
- MRQA	-	58.59(↓7.07)	73.37(↓4.86)
- CQA	-	63.64(↓2.01)	76.55(↓1.68)
- FT. on Doc2Dial	-	65.66(↓0.00)	76.67(↓1.56)

Results on the testdev set

Model	EM	F1
$QA4KGD_{KI}$	**64.42**	**77.27**
- Post-proc.	64.29(↓0.13)	77.27(↓0.00)
- MRQA	59.09(↓5.33)	73.00(↓4.27)
- CQA	62.39(↓2.03)	75.83(↓1.44)
- FT. on Doc2Dial	62.52(↓1.90)	75.44(↓1.83)

Results on the test set

- It is not feasible to rely on domain-specific data with limited high-quality resources.
  - *Except post-processing, task adaptation on Doc2Dial dataset shows the least effect to the performance of QA4KGD.*

- Pre-training on general QA dataset effectively help QA4KGD to obtain knowledge identification skills.
  - *Pre-training on MRQA datasets affects the model performance the most and improves both automatic metrics significantly.*

PPL: Exponent of negative likelihood of the target response;
F1: Unigram overlap between the generated and target responses.

Analysis: Ablation Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Response Generation

Model	SacreBLEU		
	Valid	Testdev	Test
Without Post-Proc.			
QA4KGD	**39.72**	**36.61**	**38.72**
- WoW pre-training	16.87(↓22.85)	16.75(↓19.86)	17.74(↓20.98)
- Doc2Dial (Zero-shot)	23.64(↓16.08)	21.22(↓15.39)	22.00(↓16.72)
- KI module + RoBERTa$_{large}$	40.04(↑0.32)	36.52(↓0.09)	37.91(↓0.81)

- By adapting to Doc2Dial dataset, QA4KGD generate more relevant responses to better resolve users' inquiries.

	Fluency			Relevance		
	Win	Tie	Lose	Win	Tie	Lose
w/ vs. *w/o* wow pre-training	**55.33%**	26.00%	18.67%	25.33%	38.00%	**36.67%**
w/ vs. *w/o* Doc2Dial	26.67%	57.33%	16.00%	**23.33%**	62.67%	14.00%
w/ KI module vs. *w/* RoBERTa	10.00%	86.67%	3.33%	10.67%	84.00%	5.33%

	Faithfulness
QA4KGD	3.79
w/o wow pre-training	**3.49**
w/o Doc2Dial	3.73
w/o KI module *w/* RoBERTa$_{large}$	3.70

Fluency: Whether the generated responses are complete, grammatically correct, and self-consistent without repetition;
Relevance: Whether the responses are relevant to the dialogue history and bettervresolve the user's inquiry;
Faithfulness: A faithful response should be fully supported by the dialogue context and correctly convey the information in external knowledge;
Results in bold: Statistically significant based on a pair-wise individual t-test ($p<0.05$).

Analysis: Ablation Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Response Generation

Model	SacreBLEU		
	Valid	Testdev	Test
Without Post-Proc.			
QA4KGD	**39.72**	**36.61**	**38.72**
- WoW pre-training	16.87(↓22.85)	16.75(↓19.86)	17.74(↓20.98)
- Doc2Dial (Zero-shot)	23.64(↓16.08)	21.22(↓15.39)	22.00(↓16.72)
- KI module + RoBERTa$_{\text{large}}$	40.04(↑0.32)	36.52(↓0.09)	37.91(↓0.81)

- Pre-training on the WoW dataset significantly improves the fluency and faithfulness of the generated responses.
- QA4KGD without pre-training tends to generate more relevant responses.

	Fluency			Relevance		
	Win	Tie	Lose	Win	Tie	Lose
w/ vs. *w/o* wow pre-training	**55.33%**	26.00%	18.67%	25.33%	38.00%	**36.67%**
w/ vs. *w/o* Doc2Dial	26.67%	57.33%	16.00%	**23.33%**	62.67%	14.00%
w/ KI module vs. *w/* RoBERTa	10.00%	86.67%	3.33%	10.67%	84.00%	5.33%

	Faithfulness
QA4KGD	3.79
w/o wow pre-training	**3.49**
w/o Doc2Dial	3.73
w/o KI module *w/* RoBERTa$_{\text{large}}$	3.70

Fluency: Whether the generated responses are complete, grammatically correct, and self-consistent without repetition;
Relevance: Whether the responses are relevant to the dialogue history and bettervresolve the user's inquiry;
Faithfulness: A faithful response should be fully supported by the dialogue context and correctly convey the information in external knowledge;
Results in bold: Statistically significant based on a pair-wise individual t-test ($p<0.05$).

Analysis: Case Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Responses from QA4KGD w/o pre-training are not trustworthy even though they seem to be more relevant.

	Model/Party	Utterance
Dialogue History	User	And what about the other requirements?
	System	Did you NOT receive a dishonorable discharge?
	User	No, I actually received a dishonorable discharge. That wasn't my fault though. Jim screwed up and our whole group was disciplined for it, we did nothing wrong!
Identified Knowledge	KI module	you may be eligible for the Veterans Pension program.
	RoBERTa	you may be eligible for the Veterans Pension program.
Response	QA4KGD	Then you may be eligible for the veterans pension program.
	w/o WoW	In that case you may not be eligible for the veterans pension program at all. I'm afraid that you will need to apply for a COE, but we can help you find out if you are eligible for that program, so you can get your COE with us.
	w/o Doc2Dial	You may be eligible for the veterans pension program.
	w/ RoBERTa	Then you may be eligible for the veterans pension program.

Responses from QA4KGD w/o pre-training are *not faithful* to the external knowledge.

Knowledge misuse
Hallucination

Analysis: Case Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

The noise within documents requires task-specific fine-tuning.

	Model/Party	Utterance
Dialogue History	User	What are the CDC protect and prepare communities?
	System	Learn how CDC is aggressively responding to the global outbreak of COVID-19 and preparing for the potential of community spread in the U.S.
	User	OK thank you very much
Identified Knowledge	KI module	8.5 11 English pdf icon Spanish pdf icon Chinese pdf icon Korean pdf icon Vietnamese pdf icon
	RoBERTa	Learn how CDC is aggressively responding to the global outbreak of COVID-19 and preparing for the potential of community spread in the U.S. 8.5 11 English pdf icon Spanish pdf icon
Response	QA4KGD	You are welcome.
	w/o WoW	8.5 11 English pdf icon Spanish pdf icon Chinese pdf icon Korean pdf icon Vietnamese pdf icon and other pdf icons are available for you to use. we will gladly accept your application for a COVID-19 ID #1 - ID 1.511 ID 1.5 11
	w/o Doc2Dial	You are welcome. You can also use the 11 English pdf icon to help spread the word.
	w/ RoBERTa	You are welcome.

The documents from Internet pages are *noisy*.

Models need to acquire an additional ability to *filter out noisy information* for knowledge grounding.

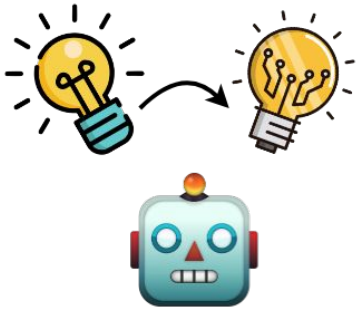
Knowledge misuse
Hallucination

Summary: Transferability

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Unstructured knowledge grounding for **transferable** open-domain dialogue systems

- *Modularity* and task-wise *pre-training*
- Simple yet *effective*
- *Domain-agnostic* with improvements by a large margin



Knowledge Grounding for Faithful Open-Domain Dialogue Systems

"Diverse and Faithful Knowledge-Grounded Dialogue Generation via Sequential Posterior Inference", Yan Xu, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, Ying Nian Wu, ICML2023

Faithfulness: Motivations

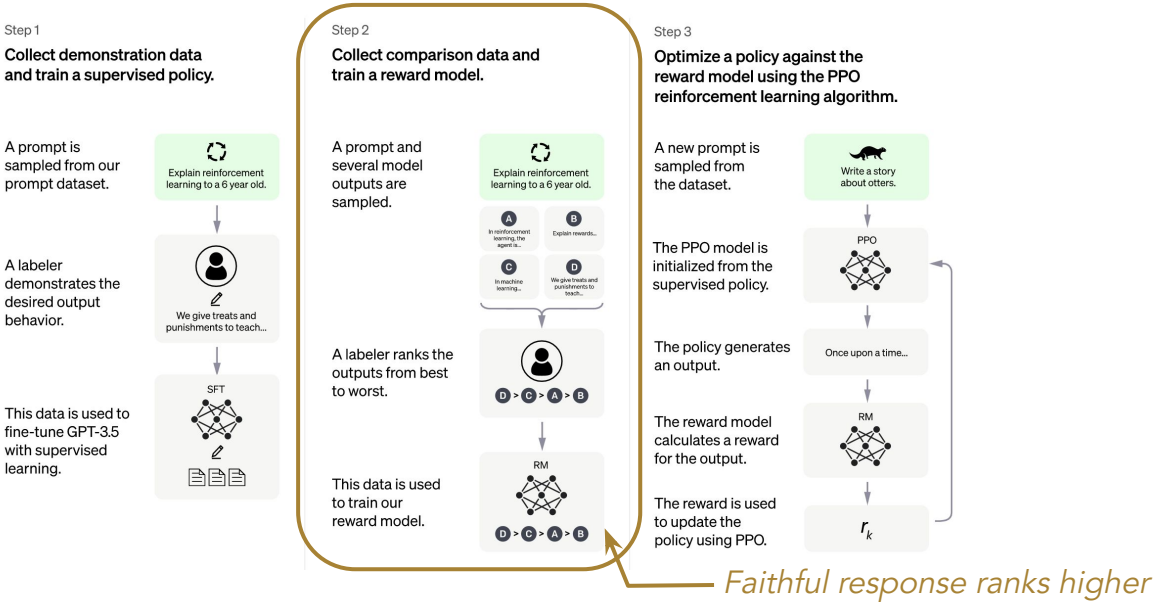
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- The capability to generate informative responses with diversity and faithfulness using factual knowledge is paramount for creating a human-like, trustworthy open-domain dialogue system.

- Prior approaches for improving the diversity of dialogue responses focus on preventing them from being dull and repetitive.
- However, optimizing for diversity alone tends to encourage the dialogue system to hallucinate non-factual responses.
- Existing explorations on faithfulness in KGD rely on additional annotations.

Introduction of Existing Work

- ChatGPT & GPT-4 incorporate a *reward model* to improve the faithfulness of generated responses. However, this methodology is notably *resource-intensive*.



[1] "GPT-4 Technical Report." OpenAI, 2023.

Our Path to Solution

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

We propose

A parameter-efficient end-to-end approach to enhance the *faithfulness* of responses to external unstructured knowledge

- Without sacrificing *diversity*;
- No additional annotation is needed.

How to enhance faithfulness?

There is no direct objective to optimize *faithfulness*.

We investigate methods:

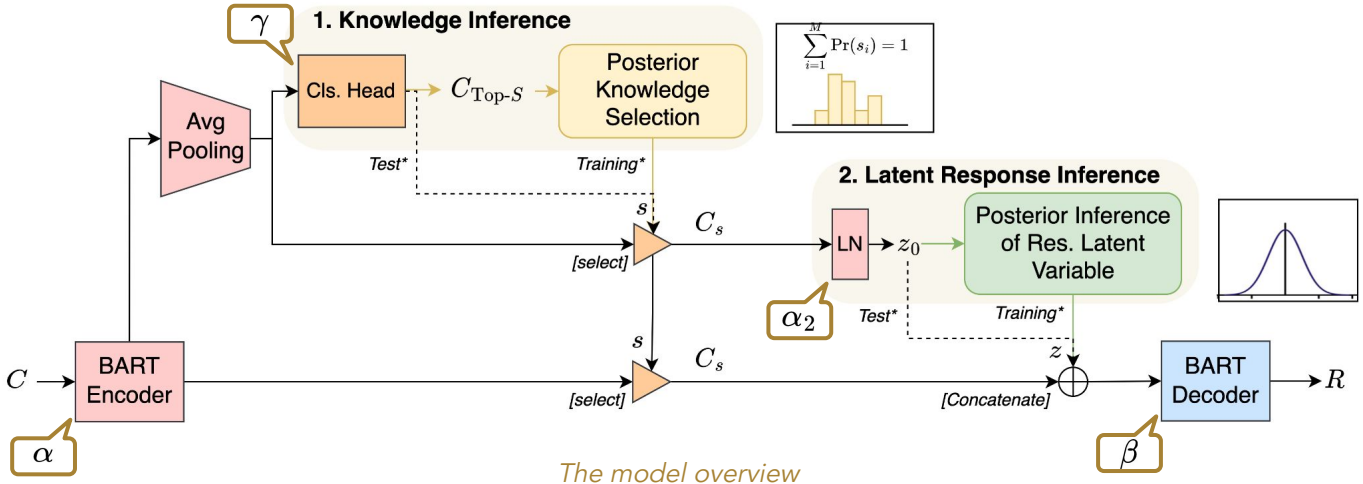
1. Enhance the *inherent correlation* between knowledge selection and response generation;
2. Provide the decoder with an explicit *high-level abstraction* of the future response.

Our Model

- **Dual latent variables**:
  - a discrete latent variable s for knowledge selection
    - s denotes the index of selected knowledge candidate
  - a continuous latent variable z for response generation
    - z can be considered as a special token or a trainable control code

$C^n = (H^n, \mathbf{K}^n)$

H^n : the dialogue history
$\mathbf{K}^n$: M knowledge sentences as candidates



The model overview

SPI: Method

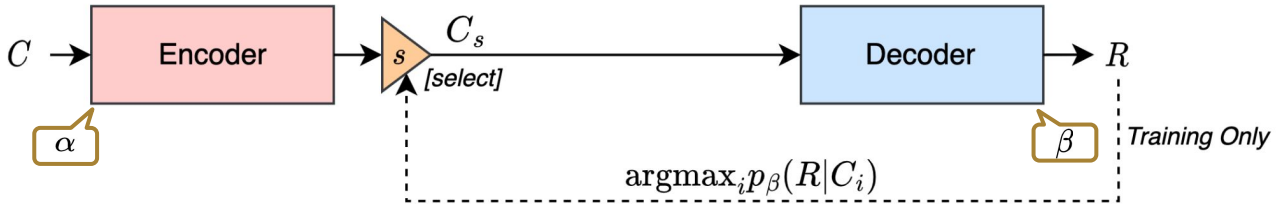
We propose **Sequential Posterior Inference (SPI)**

$$\nabla_\theta \log p_\theta(R|C) = \mathbb{E}_{p_\theta(s,z|R,C)}[\nabla_\theta \log p_\theta(s,z,R|C)]$$

- **Posterior knowledge selection** $p_\theta(s|R,C)$
  - *Select knowledge*

$C^n = (H^n, \mathbf{K}^n)$

H^n : the dialogue history
$\mathbf{K}^n$: M knowledge sentences as candidates



SPI: Method

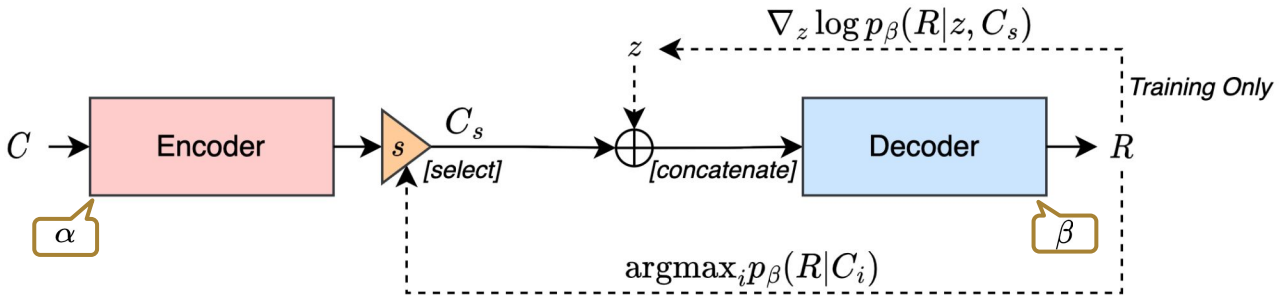
We propose **Sequential Posterior Inference (SPI)**

$$\nabla_\theta \log p_\theta(R|C) = \mathbb{E}_{p_\theta(s,z|R,C)}[\nabla_\theta \log p_\theta(s,z,R|C)]$$

$C^n = (H^n, \mathbf{K}^n)$

H^n : the dialogue history
$\mathbf{K}^n$: M knowledge sentences as candidates

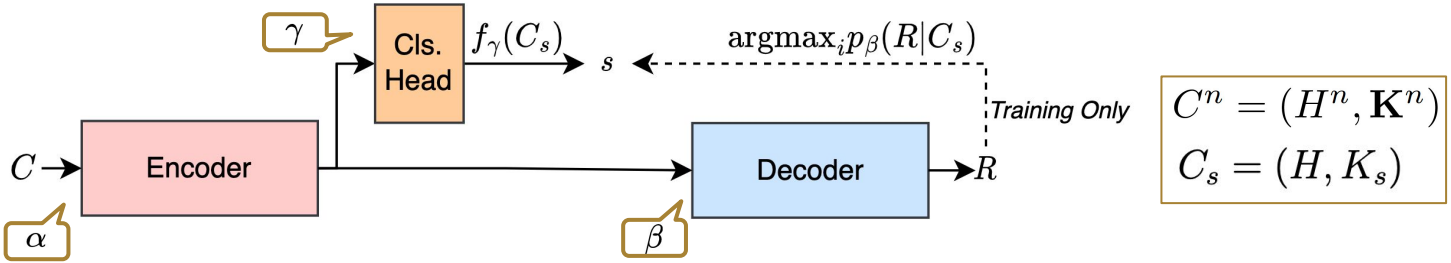
- **Posterior knowledge selection** $p_\theta(s|R,C)$
  - *Select knowledge*
- **Posterior inference of response latent variable** $p_\theta(z|R,C_s)$
  - *Infer the response latent variable*



SPI: Posterior Knowledge Selection

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior knowledge selection enhances the **inherent correlation** between knowledge selection and response generation.*



- **Uniform prior** $p(s|C) \sim U\{0, M\text{-}1\}$

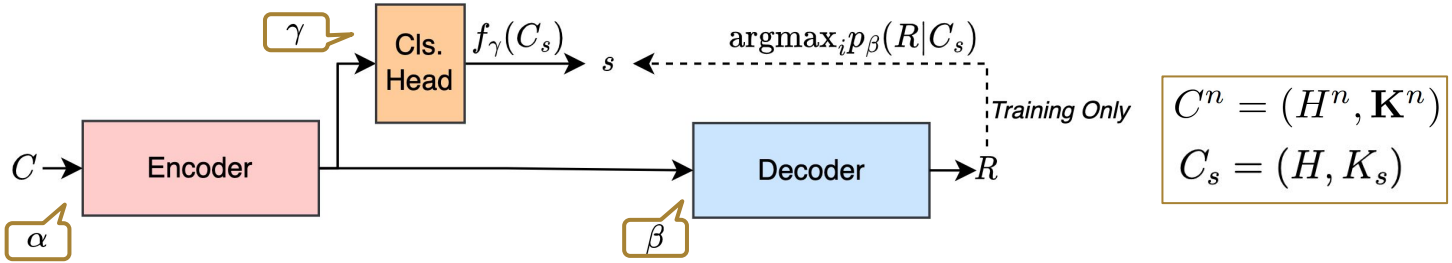
knowledge index

$$p_\theta(s|R,C) \propto p(s|C)p_\beta(R|C_s)$$
$$\propto p_\beta(R|C_s).$$

SPI: Posterior Knowledge Selection

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior knowledge selection enhances the **inherent correlation** between knowledge selection and response generation.*



- **Uniform prior** $p(s|C) \sim U\{0, M\text{-}1\}$

$$p_\theta(s|R,C) \propto p(s|C)p_\beta(R|C_s)$$
$$\propto p_\beta(R|C_s).$$

knowledge index

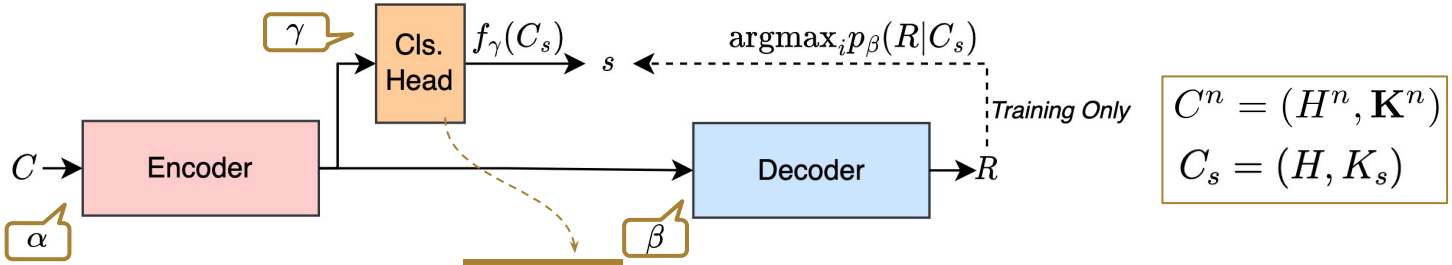
$$s = \arg\max_i p_\beta(R|C_i), i \in \{1,\ldots,M\}$$

Knowledge selection is fully dependent on the likelihood of the decoder to generate the target response with each knowledge candidate. β

SPI: Posterior Knowledge Selection

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior knowledge selection enhances the **inherent correlation** between knowledge selection and response generation.*



- Uniform prior $p(s|C) \sim U\{0, M\text{-}1\}$ with **initializer** $f_\gamma(C)$

$$p_\theta(s|R,C) \propto p(s|C)p_\beta(R|C_s)$$
$$\propto p_\beta(R|C_s).$$
$$s = \arg\max_i p_\beta(R|C_i), i \in \{1,\ldots,S\}$$

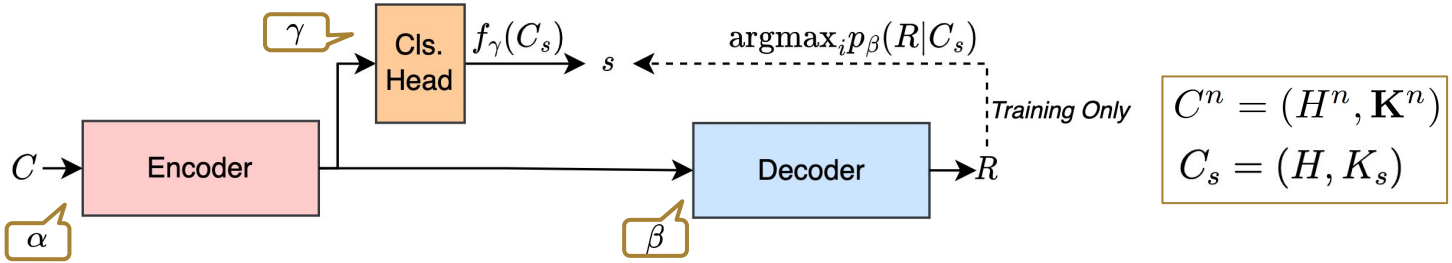
knowledge index

$S \ll M$

SPI: Posterior Knowledge Selection

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior knowledge selection enhances the **inherent correlation** between knowledge selection and response generation.*



- Uniform prior $p(s|C) \sim U\{0, M\text{-}1\}$ with initializer $f_\gamma(C)$

$$p_\theta(s|R,C) \propto p(s|C)p_\beta(R|C_s)$$
$$\propto p_\beta(R|C_s).$$
$$s = \arg\max_i p_\beta(R|C_i), i \in \{1,\ldots,S\}$$

knowledge index

$S \ll M$

- **Initializer** learns from the **posterior selection**

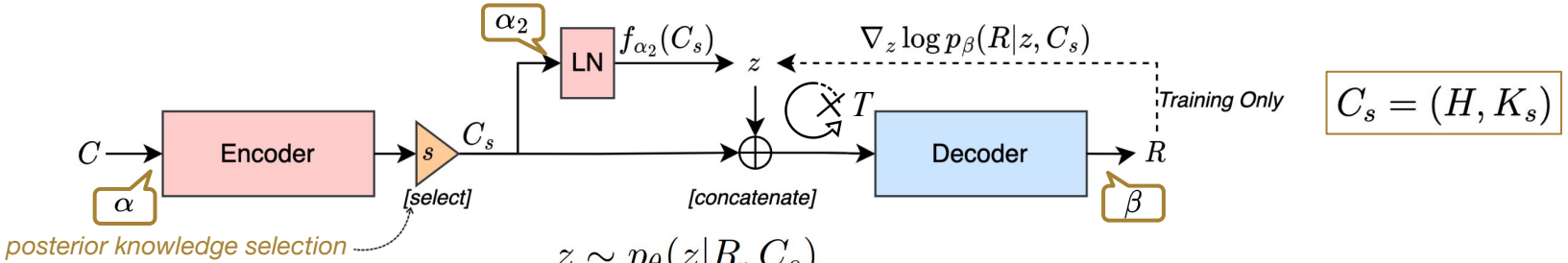
$$\mathcal{L}_{\text{CE}}(y,C) = -\sum_{i=1}^{M} y_i \log f_\gamma(C_i) + (1-y_i)\log(1-f_\gamma(C_i))$$

where ground-truth $y_i^n = 1$ if it's *gold annotation* or it's selected by *posterior knowledge selection*.

SPI: Posterior Inference of Response Latent Variable

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior inference of response latent variable provides the decoder with an **high-level abstraction** of the future response.*



- **Learning with Langevin Dynamics & short-run MCMC** [14, 15]

Initial State $z^0 \sim p_{\alpha_2}(z|C_s),$

Langevin Dynamics $z^{t+1} = z^t + \delta \nabla_z \log p_\theta(z^t|R, C_s) + \sqrt{2\delta}\epsilon_t \qquad t = 1, \ldots, T$

MCMC: Monte Carlo Markov Chain

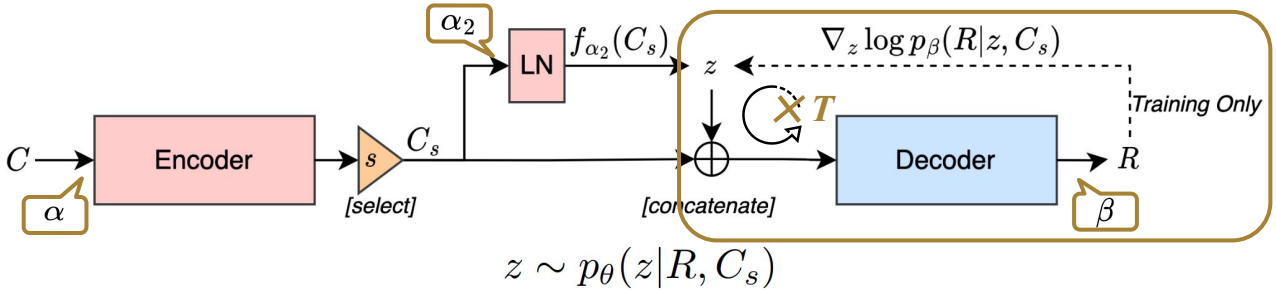
[14] "Learning Multi-layer Latent Variable Model via Variational Optimization of Short Run MCMC for Approximate Inference". Nijkamp, Erik, et al. ECCV 2020.

[15] "Generative Text Modeling through Short Run Inference." Pang, Bo, et al. EACL 2021.

SPI: Posterior Inference of Response Latent Variable

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior inference of response latent variable provides the decoder with an **high-level abstraction** of the future response.*



$$C_s = (H, K_s)$$

$$z \sim p_\theta(z|R, C_s)$$

- **Learning with Langevin Dynamics & short-run MCMC** [14, 15]

Initial State $z^0 \sim p_{\alpha_2}(z|C_s),$

Langevin Dynamics $z^{t+1} = z^t + \delta \nabla_z \log p_\theta(z^t|R, C_s) + \sqrt{2\delta}\epsilon_t$

$t = 1, \ldots, T$ hyper-parameter

MCMC: Monte Carlo Markov Chain

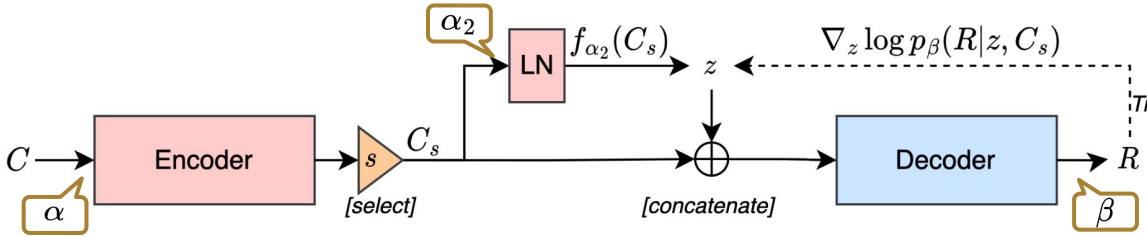
[14] "Learning Multi-layer Latent Variable Model via Variational Optimization of Short Run MCMC for Approximate Inference". Nijkamp, Erik, et al. ECCV 2020.

[15] "Generative Text Modeling through Short Run Inference." Pang, Bo, et al. EACL 2021.

SPI: Posterior Inference of Response Latent Variable

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior inference of response latent variable provides the decoder with an **high-level abstraction** of the future response.*

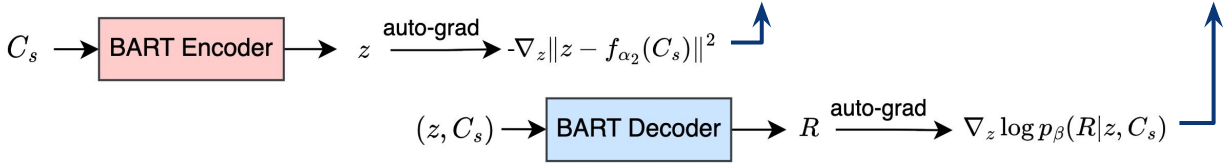


$$z \sim p_\theta(z|R, C_s)$$

Langevin Dynamics $\quad z^{t+1} = z^t + \delta \nabla_z \log p_\theta(z^t|R, C_s) + \sqrt{2\delta}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$

where the gradient is tractable $\quad \nabla_z \log p_\theta(z|R, C_s) = \nabla_z \log p_{\alpha_2}(z|C_s) + \nabla_z \log p_\beta(R|z, C_s)$

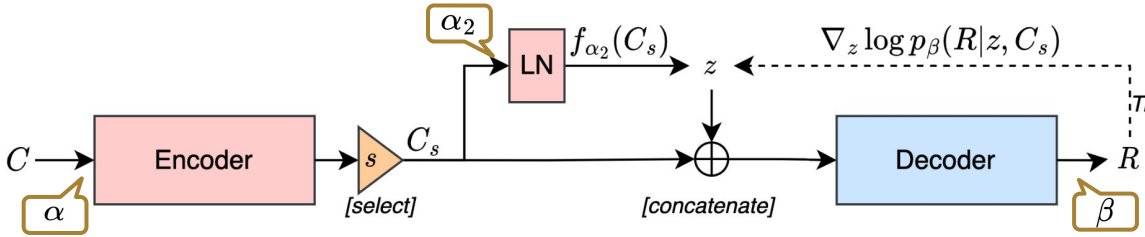
Gaussian Distribution **Response Generation**



SPI: Posterior Inference of Response Latent Variable

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

*Posterior inference of response latent variable provides the decoder with an **high-level abstraction** of the future response.*



$C_s = (H, K_s)$

$$z \sim p_\theta(z|R, C_s)$$

Langevin Dynamics $$z^{t+1} = z^t + \delta \nabla_z \log p_\theta(z^t|R, C_s) + \sqrt{2\delta}\epsilon_t \quad \text{, where } \epsilon_t \sim \mathcal{N}(0, \mathbf{I})$$

where the gradient is tractable $$\nabla_z \log p_\theta(z|R, C_s) = \nabla_z \log p_{\alpha_2}(z|C_s) + \nabla_z \log p_\beta(R|z, C_s)$$

Gaussian Distribution **Response Generation**

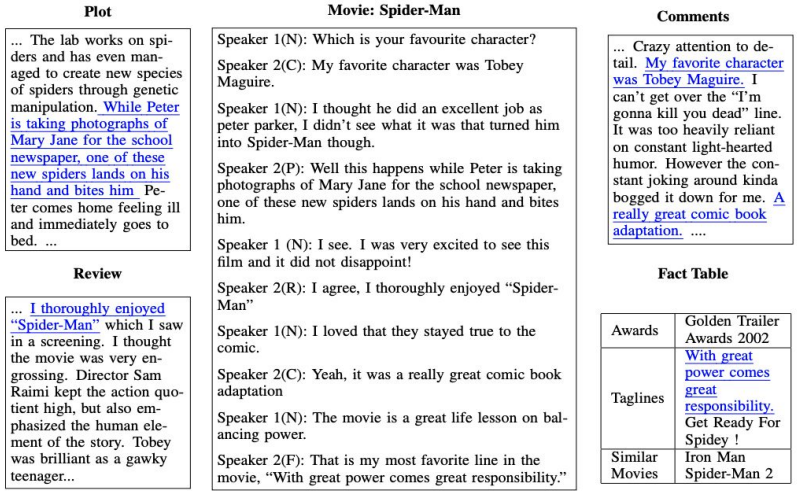
The prior distribution of z follows a Gaussian Distribution

z is optimized to maximize the log-likelihood of the decoder to generate the target response given dialogue context

Experiments

Datasets

- Wizard of Wikipedia (WoW) [3] is a KGD dataset across 1,365 topics
  - 18,430/1,948/965/968 train/valid/test seen/test unseen dialogues
  - Ground on Wikipedia articles
- Holl-E [16] is a KGD dataset in the movie domain
  - 7,228/930/913 train/valid/test dialogues
  - Ground on Wikipedia, reviews, and etc.



Plot

... The lab works on spiders and has even managed to create new species of spiders through genetic manipulation. While Peter is taking photographs of Mary Jane for the school newspaper, one of these new spiders lands on his hand and bites him. Peter comes home feeling ill and immediately goes to bed. ...

Review

... I thoroughly enjoyed "Spider-Man" which I saw in a screening. I thought the movie was very engrossing. Director Sam Raimi kept the action quotient high, but also emphasized the human element of the story. Tobey was brilliant as a gawky teenager...

Movie: Spider-Man

Speaker 1(N): Which is your favourite character?

Speaker 2(C): My favorite character was Tobey Maguire.

Speaker 1(N): I thought he did an excellent job as peter parker, I didn't see what it was that turned him into Spider-Man though.

Speaker 2(P): Well this happens while Peter is taking photographs of Mary Jane for the school newspaper, one of these new spiders lands on his hand and bites him.

Speaker 1 (N): I see. I was very excited to see this film and it did not disappoint!

Speaker 2(R): I agree, I thoroughly enjoyed "Spider-Man"

Speaker 1(N): I loved that they stayed true to the comic.

Speaker 2(C): Yeah, it was a really great comic book adaptation

Speaker 1(N): The movie is a great life lesson on balancing power.

Speaker 2(F): That is my most favorite line in the movie, "With great power comes great responsibility."

Comments

... Crazy attention to detail. My favorite character was Tobey Maguire. I can't get over the "I'm gonna kill you dead" line. It was too heavily reliant on constant light-hearted humor. However the constant joking around kinda bogged it down for me. A really great comic book adaptation. ....

Fact Table

Awards	Golden Trailer Awards 2002
Taglines	With great power comes great responsibility. Get Ready For Spidey !
Similar Movies	Iron Man Spider-Man 2

An example from Holl-E dataset

[16] "Towards Exploiting Background Knowledge for Building Conversation Systems". Moghe, Nikita, et al. EMNLP 2018.

Results: Automatic Evaluation

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

	Model	WoW Seen								WoW Unseen							
		PPL↓	B3↑	B4↑	R1↑	R2↑	Dist-1↑	Dist-2↑	Acc↑	PPL↓	B3↑	B4↑	R1↑	R2↑	Dist-1↑	Dist-2↑	Acc↑
Latent-variable based	$BART_{SKT}$	20.3	7.6	4.4	19.4	5.4	6.8	30.3	26.8	22.3	–	4.6	19	4.7	5.2	24.5	18.3
	ZRKGC	40.4	2.8	1.8	18.6	2.4	5.4	22.5	–	41.5	18.6	1.1	18.5	2.4	3.4	15.6	–
	PIPM	42.7	–	3.3	19.9	7.3	–	26.4	27.7	65.7	–	2.5	17.6	5.4	–	17.7	19.4
	CoLV	39.6	–	2.9	20.6	7.9	–	29.7	30.1	54.3	–	2.1	19.7	6.3	–	20.1	18.9
Others	$BART_{cat}$	19.7	6.7	4.3	19.3	5.1	7.1	29.9	–	24.5	–	4.1	18.9	4.5	5.3	22.2	–
	$BART_{FiD}$	**9.5**	7.9	5.8	20.9	7.8	10.4	39.6	–	**10.5**	8.1	6.1	20.9	7.9	6.7	24.2	–
	DRD	23.0	7.5	5.5	18.0	–	–	–	–	25.6	16.5	4.3	16.5	–	–	–	–
	KAT-TSLF	14.4	9.1	6.7	21.7	7.6	9.5	38.3	–	15.8	8.3	6.0	20.7	7.2	6.7	**26.0**	–
	KnowledGPT	19.2	9.5	7.2	22.0	7.9	8.9	36.2	28.0	22.3	8.3	6.0	20.5	6.7	6.0	23.8	24.0
Ours	**SPI**	17.1	**10.2**	**7.7**	**22.7**	**8.8**	**10.8**	**40.9**	36.2	19.1	9.6	7.3	**22.0**	**8.5**	**6.9**	24.3	34.6

Automatic results on WoW comparing with various baseline models

	Model	PPL↓	B4	R1	R2	Dist-2	Acc
Latent-variable based	SKT	48.9	-	29.8	23.1	-	29.2
	DukeNet	42.7	19.2	32.6	19.6	28.5	30.4
	PIPM	39.2	18.3	30.8	24.0	27.2	30.7
	CoLV	34.8	20.3	32.0	25.8	29.9	32.7
Ours	**SPI**	**12.6**	**30.7**	**38.3**	**31.7**	**30.6**	**38.3**

Automatic results on Holl-E

- Our proposed model, SPI, achieves SOTA performance on both WoW and Holl-E datasets.

B3/B4/R1/R2: BLEU-3/BLEU-4/ROUGE-1/ROUGE-2;
Acc: Knowledge selection accuracy.

Results: Diversity and Faithfulness

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Model	WoW Seen		WoW Unseen	
	Dist-1↑	Dist-2↑	Dist-1↑	Dist-2↑
$BART_{cat}$	7.1	29.9	5.3	22.2
$BART_{SKT}$	6.8	30.3	5.2	24.5
$BART_{FiD}$	10.4	39.6	6.7	24.2
ZRKGC	5.4	22.5	3.4	15.6
DRD	–	–	–	–
PIPM	–	26.4	–	17.7
CoLV	–	29.7	–	20.1
KAT-TSLF	9.5	38.3	6.7	**26.0**
KnowledGPT	8.9	36.2	6.0	23.8
SPI	**10.8**	**40.9**	**6.9**	24.3

Automatic diversity metrics

Model	Oracle Performance					FeQA	QuestEval	
	PPL↓	B3	B4	R1	R2		RD	RF
WoW Seen								
KnowledGPT	9.1	19.2	15.5	34.5	17.3	48.1	42.2	43.5
SPI	**8.7**	**20.0**	**16.3**	**36.1**	**18.7**	**49.2**	44.4	46.0
WoW Unseen								
KnowledGPT	9.8	18.3	14.6	33.8	16.5	47.4	41.0	42.2
SPI	**9.2**	**20.1**	**16.3**	**36.0**	**18.7**	**49.6**	44.0	**45.7**

Automatic faithfulness metrics

- SPI effectively improves both diversity and faithfulness of the generated responses.

Model	Fluency		Relevance		Faithfulness	
	Seen	Un.	Seen	Un.	Seen	Un.
KnowledGPT	62.5%	60.3%	70.8%	62.2%	3.33	3.42
SPI	**88.7%**	**83.3%**	79.8%	**74.4%**	**3.66**	**3.65**

Human evaluation on faithfulness and etc.

FeQA: Faithfulness measure based on a QG-QA based framework;
QuestEval: Similar to FeQA, but with reference-free (RF) and reference-dependent (RD) settings.
Results in bold in human evaluation: Statistically significant based on a pair-wise individual t-test ($p<0.05$).

Analysis: Ablation Study

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

- **Both** posterior knowledge selection and posterior inference of response latent variable **contribute** to the improvement of faithfulness and diversity.
- Posterior inference of z brings more improvements to the **unseen** domain.

Top-S	WoW Seen						WoW Unseen					
	B-4	R-2	Dist-2	FeQA	Q.E.(RD/RF)	Acc	B-4	R-2	Dist-2	FeQA	Q.E.(RD/RF)	Acc
1	7.3	8.4	36.6	40.4	41.1/43.0	**37.0**	6.9	7.7	22.5	39.2	39.9/41.8	**34.7**
3	7.4	8.3	39.4	40.7	41.4/43.2	34.1	7.0	7.8	22.5	40.5	40.5/42.2	32.2
5 (Ours)	**7.7**	**8.8**	40.9	**49.2**	**44.4/46.0**	36.2	**7.3**	**8.5**	24.3	**49.6**	**44.0/45.7**	34.6
10	7.2	8.8	**41.1**	48.0	42.4/44.2	36.4	7.3	8.4	**24.4**	47.7	42.3/44.0	34.6

Impact of Top-S selection

Langevin Steps	WoW Seen					WoW Unseen					Tr. Time (/Epoch)
	B4	R2	Dist-2	FeQA	Q.E.(RD/RF)	B4	R2	Dist-2	FeQA	Q.E.(RD/RF)	
0	7.4	8.7	40.3	47.4	43.8/45.6	6.9	8.2	23.5	48.0	42.9/44.6	3.50hrs
1	7.6	8.7	40.3	47.9	44.2/45.9	**7.4**	8.4	23.1	47.9	43.5/45.1	3.56hrs
5 (Ours)	**7.7**	**8.8**	**40.9**	**49.2**	**44.4/46.0**	7.3	**8.5**	**24.3**	**49.6**	**44.0/45.7**	3.68hrs

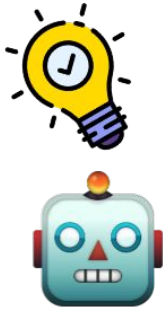
Impact of the number of Langevin steps (T)

Summary: Faithfulness

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Unstructured knowledge grounding for **faithful** open-domain dialogue systems

- Novel and effective
  - Enhancing the inherent correlation with knowledge selection and response generation
  - Providing the decoder with a high-level abstraction of the future response
- *Alignment* between knowledge selection and response generation, thus improve knowledge utilization ability
- Can also contribute to *LLM-based dialogue models.*



Conclusion

Conclusion

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Goal: Building knowledgeable and trustworthy open-domain dialogue models with unstructured knowledge grounding

Knowledgeable

- We are the first to adapt *to new topics* with lightweight adapters and pseudo-conversational style pre-training to build a more knowledgeable open-domain dialogue system.

Trustworthy

- We *decompose* the KGD task into sub-modules and conduct task-wise pre-training with multi-task learning to acquire transferable knowledge grounding skills and ensure trustworthy performance in *niche domains*.
- We propose a probabilistic model with dual latent variable and learning algorithm SPI to enhance the faithfulness of responses *without additional annotations*.

Thesis Contributions

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Scientific Contributions

- We are the first to point out that faithfulness can be enhanced by aligning knowledge selection and response generation to improve the knowledge utilization. We make two hypotheses that
  (1) enhancing the inherent correlation between knowledge selection and response generation,
  (2) providing the decoder (generator) with a high-level abstraction.
  Our experiments have validated these hypotheses.

Methodological Contributions

- We *pioneer* the research direction to injecting new unstructured knowledge into PLMs with lightweight adapters through *conversational-style pre-training*.
- We are the first to inject posterior information of response generation by querying the decoder and show its effectiveness.

Empirical Contributions

- We are among the first to propose *task decomposition* to improve the transferability of knowedge grounding skills in niche domains.
- The adapter can *store certain knowledge* and further benefit response generation.

Publications

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Total 20 publications at following venues: ACL, ICML, EMNLP, AAAI, ACM Computing, AACL, LREC, WMT, and workshops

1. **Yan Xu**, Deqian Kong, Dehong Xu, Ziwei Ji, Bo Pang, Pascale Fung, Ying Nian Wu. "Diverse and Faithful Knowledge-Grounded Dialogue Generation via Sequential Posterior Inference." ICML 2023.
2. **Yan Xu**, Mahdi Namazifar, Devamanyu Hazarika, Aishwarya Padmakumar, Yang Liu, Dilek Hakkani-Tür. "KILM: Knowledge Injection into Encoder-Decoder Language Models." ACL 2023.
3. **Yan Xu**, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan Su, Pascale Fung."Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters." ACL 2022 DialDoc.
4. **Yan Xu**, Etsuko Ishii, Genta Indra Winata, Zhaojiang Lin, Andrea Madotto, Zihan Liu, Peng Xu, Pascale Fung. "CAiRE in DialDoc21: Data Augmentation for Information Seeking Dialogue System." ACL 2021 DialDoc.
5. Dan Su*, **Yan Xu***, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. "Generalizing Question Answering System with Pre-trained Language Model Fine-tuning." EMNLP 2019 MRQA.
6. Zihan Liu*, **Yan Xu***, Genta Indra Winata, and Pascale Fung. "Incorporating Word and Subword Units in Unsupervised Machine Translation Using Language Model Rescoring." WMT 2019.
7. Etsuko Ishii*, Bryan Wilie*, **Yan Xu***, Samuel Cahyawijaya, Pascale Fung. "Integrating Question Rewrites in Conversational Question Answering: A Reinforcement Learning Approach." ACL 2022 SRW.
8. Bryan Wilie, **Yan Xu**, Willy Chung, Samuel Cahyawijaya, Holy Lovenia, Pascale Fung. "PICK: Polished & Informed Candidate Scoring for Knowledge-Grounded Dialogue Systems." AACL 2023.

Publications

9. Ziwei Ji, **Yan Xu**, I-Tsun Cheng, Samuel Cahyawijaya, Rita Frieske, Etsuko Ishii, Min Zeng, Andrea Madotto, Pascale Fung. "VScript: Controllable Script Generation with Visual Presentation." AACL Demo 2022.
10. Zihan Liu, **Yan Xu**, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, Pascale Fung. "CrossNER: Evaluating Cross-Domain Named Entity Recognition." AAAI 2021.
11. Dan Su, **Yan Xu**, Wenliang Dai, Ziwei Ji, Tiezheng Yu, Pascale Fung. "Multi-hop Question Generation with Graph Convolutional Network." EMNLP 2020.
12. Dan Su, **Yan Xu,** Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, Pascale Fung. "CAiRE-COVID: a question answering and query-focused multi-document summarization system for covid-19 scholarly information management." EMNLP 2020 NLP for COVID-19.
13. Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, **Yan Xu**, Pascale Fung. "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity." AACL 2023.
14. Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, **Yan Xu**, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, Pascale Fung. "Survey of hallucination in natural language generation." ACM Computing Surveys. 2023.
15. Yejin Bang, Nayeon Lee, Tiezheng Yu, Leila Khalatbari, **Yan Xu**, Dan Su, Elham J Barezi, Andrea Madotto, Hayden Kee, Pascale Fung. "Towards Answering Ethical Quandary Questions." AAAI 2023 AI for Social Good.

Publications

16. Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, **Yan Xu**, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, Qifeng Chen, Xiaojuan Ma, Bertram E Shi, Pascale Fung. "ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation." LREC 2022.
17. Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, **Yan Xu**, Zihan Liu, Zhaojiang Lin, Pascale Fung. "Learning Knowledge Bases with Parameters for Task-Oriented Dialogue Systems." EMNLP 2020.
18. Zihan Liu, Jamin Shin, **Yan Xu**, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. "Zero-shot Cross-lingual Dialogue Systems with Transferable Latent Variables.", EMNLP 2019.
19. Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Zihan Liu, **Yan Xu**, Cong Gao, and Pascale Fung. "Learning to learn sales prediction with social media sentiment." IJCAI 2019 FinTech.
20. Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Jamin Shin, **Yan Xu**, Peng Xu, Pascale Fung. "CAiRE_HKUST at SemEval-2019 Task 3: Hierarchical Attention for Dialogue Emotion Classification." NAACL 2019 SemEval.

Competitions & Awards

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Best Student Paper Award *(DialDoc@ACL 2022)*

BEST STUDENT PAPER AWARD

Is presented to

Yan Xu, Etsuko Ishii, Samuel Cahyawijaya, Zihan Liu, Genta Indra Winata, Andrea Madotto, Dan SU, Pascale Fung

Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters

For winning the Best Student Paper Award

DIALDOC@ACL 2022

May 26, 2022

DATE

S. Feng, C. Tang, S. Vakulenko, K. Wim, Z. Wu, C. Yuan

ORGANIZERS

Competitions & Awards

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

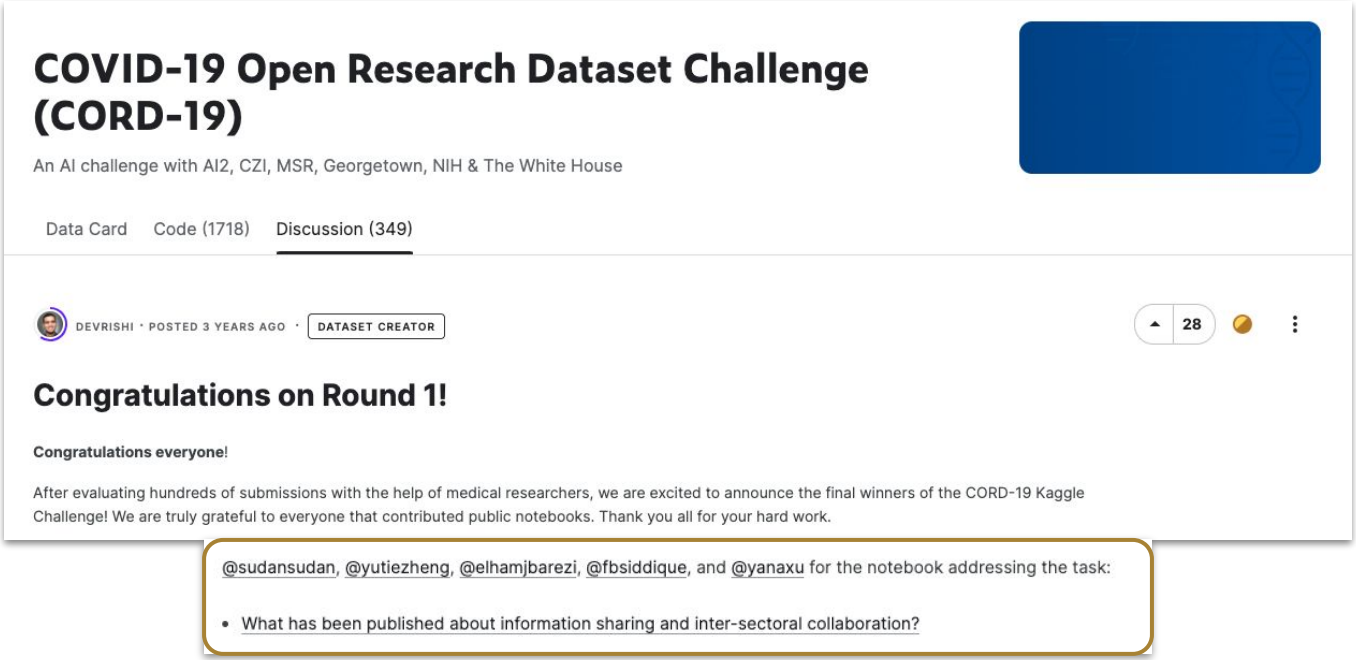
Top1 *on Chatbot Millionaire Challenge*

https://www.hkstp.org/news-room/hkstp-s-aiplusu-explore-and-experience-exhibition-brings-together-the-ai-community-to-charter-next-path-of-growth/

Competitions & Awards

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Kaggle winner *on COVID-19 Open Research Dataset Challenge (CORD-19)*

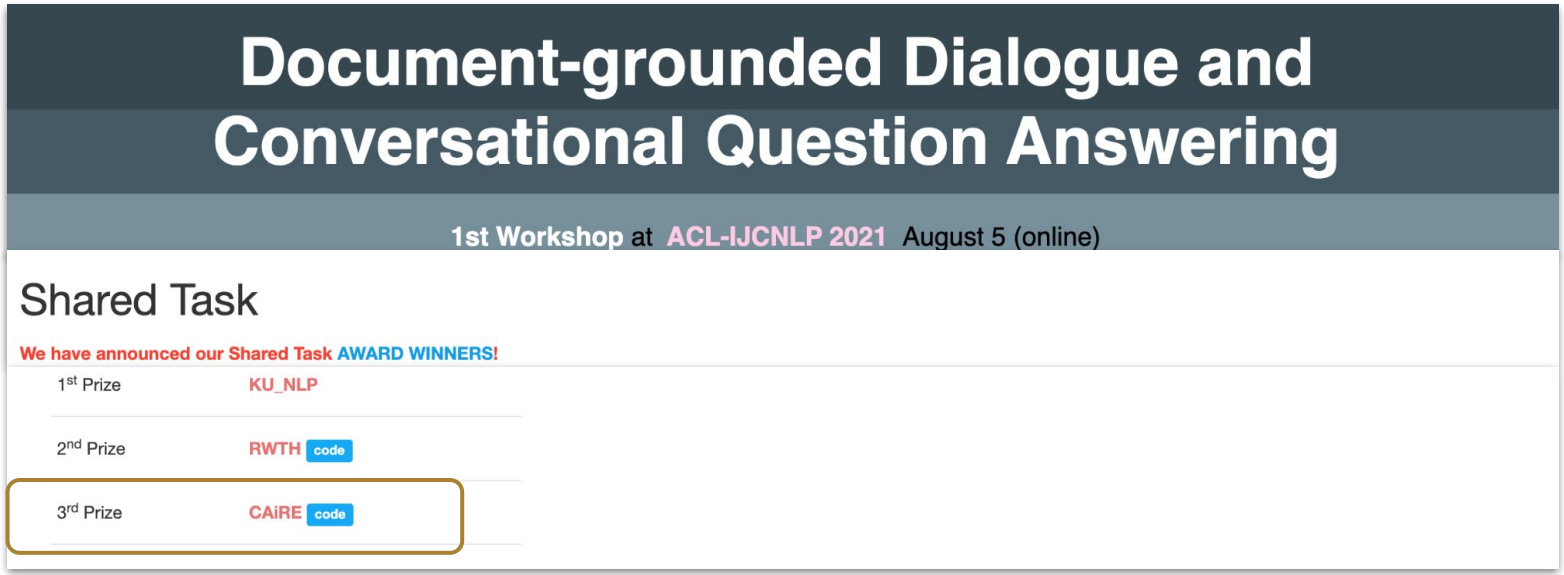


https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/148807

Competitions & Awards

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

3rd prize *on DialDoc21 Shared Task*



https://doc2dial.github.io/workshop2021/shared.html

Acknowledgement

Acknowledgement

- I would like to thank my supervisor Prof. Pascale Fung for providing guidance and motivation!
- I would like to also express my appreciation to Prof. Wei Zhang, Prof. Qifeng Chen, Prof. Dan Xu, and Prof. Tatsuya Kawahara to be on my thesis examination committee! Also many thanks to Prof. Dimitris Papadopoulos to host this thesis defense!
- I am grateful to collaborate with amazing labmates during my PhD journey!
- I would like to express my gratitude to my family for their unconditional love and support!

Thanks for Attention

Any question is welcome!