中国人民大学 RENMIN UNIVERSITY OF CHINA 高瓴人工智能学院 Gaoling School of Artificial Intelligence

Knowledge-Aware Dialogues

Rui Yan

Gaoling School of Artificial Intelligence

Renmin University of China

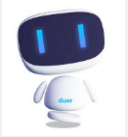
ruiyan@ruc.edu.cn

Conversational AI

Virtual Personal Assistant


Microsoft Cortana


Apple Siri


Baidu DuMi

Smart Speaker


Amazon Echo


Google Home


AliGenie

Conversational Agent (ChatBot)


Microsoft Xiaoice


Microsoft Rinna

Turing Robot

E-commerce Customer Service Robot

Alime


JingDong Jimi


AiKF

Background Info.

- Mainstream framework for dialogues
  - Retrieval methods
  - Generation methods
- Building a dialogue system has gained increasing interest
  - Industrial applications: Microsoft XiaoIce, Amazon Alexa, 小爱同学
- Limitations:
  - Only able to awkwardly catch up with the conversation
  - Can not dive into a specific topic with humans owing to the lack of knowledge of the subject

Previous Studies

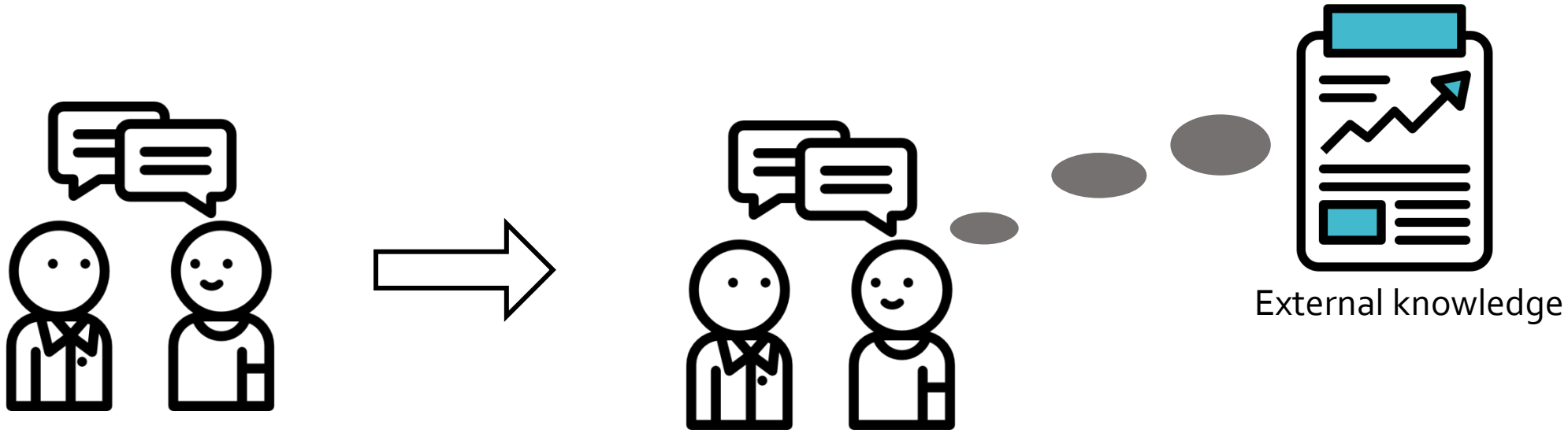
- Build benchmarks with the source of Wikipedia
  - A dataset for document grounded conversations, EMNLP 2019
  - Wizard of wikipedia: Knowledge-powered conversational agents, ICLR 2019
- Toward appropriate knowledge selection
  - Learning to Select Knowledge for Response Generation in Dialog Systems, IJCAI 2019
  - Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue, ICLR 2020
- KGC dialogue system
  - **[Zhao et al., IJCAI 2019, EMNLP 2020]**

Knowledge in Dialogues

Knowledge	
Name	The inception
Year	2009
Director	Christopher Nolan
Cast	Leonardo DiCaprio as Dom Cobb, a professional thief…
Introduction	Dominick Gobband Arthur are extractors, who perform corporate espionage using an experimental military technology to infiltrate the subconscious of theirtargets and extract valuable information through a shared dream world. ……
Conversation	
User2:	Hey have you seen the inception?
User1:	No, I have not but have heard of it. What is it about
User2:	It's about extractors that perform experiments using military technology on people to retrieve info about their targets.
User1:	Sounds interesting do you know which actors are in it?
User2:	……

Knowledge-Grounded Conversations

- From traditional dialogue setting to knowledge-grounded dialogue setting



Grounding dialogue agents with background knowledge

中国人民大学 RENMIN UNIVERSITY OF CHINA 高瓴人工智能学院 Gaoling School of Artificial Intelligence

Low-Resource Knowledge Grounded Dialogue Generation

Xueliang Zhao, Wei Wu, Chongyang Tao, Can Xu, Dongyan Zhao and **Rui Yan**

ICLR 2020

Problems

- Some recent work resorts to crowd-sourcing and builds benchmarks with the source of Wikipedia
- There is still a long way to go for application of the existing models in real scenarios
  - When they are applied to documents out of domain, their performance drops dramatically
  - It is difficult to collect enough training data for a new domain or a new language, as human effort is expensive

Challenges

- How to learn a model with as few knowledge-grounded dialogues as possible, yet the model achieves state-of-the-art performance ?
- How to make the model generalize well on out-of-domain documents ?

Methodology

- Make parameters that rely on knowledge-grounded dialogues small and independent
- Encoder
  - **Context encoder**
  - **Knowledge encoder**
- Decoder
  - **Language model**
  - **Context processor**
  - Knowledge processor
- Decoding manager

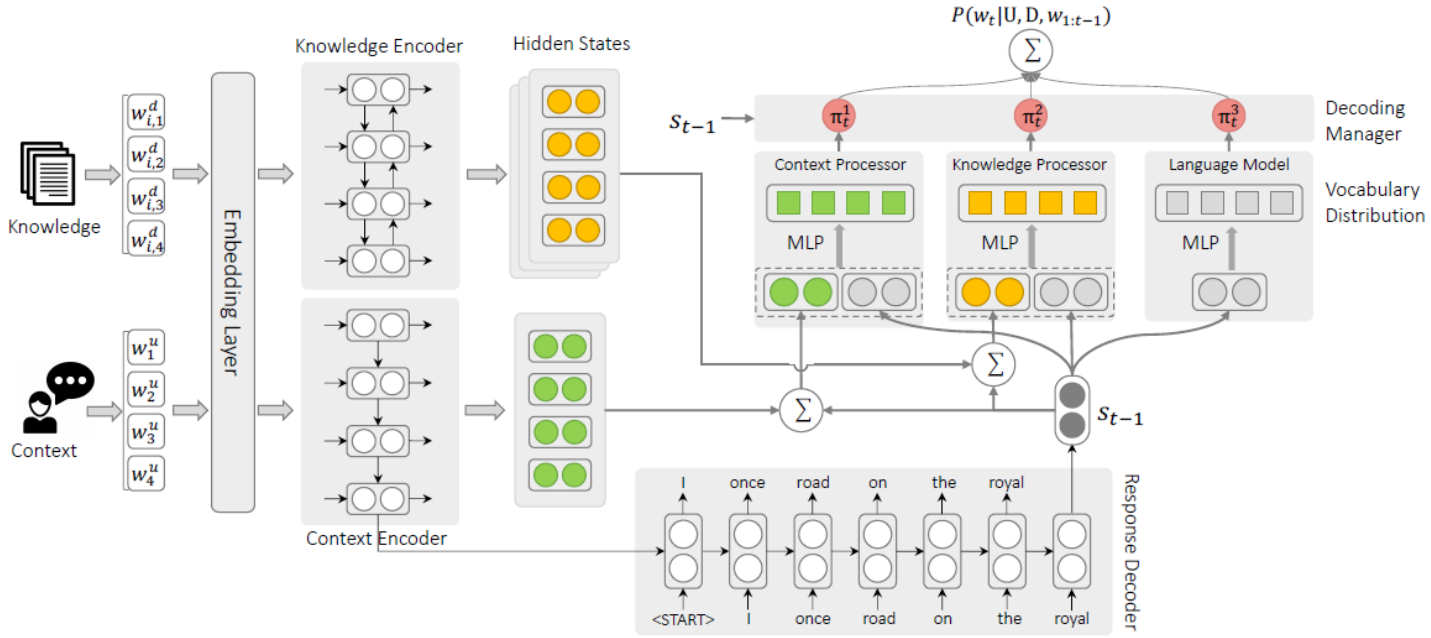


Figure 1: Architecture of the generation model.

Pre-trained with the Ungrounded Dialogues

- Language model

$$P(w_t^r | w_{1:t-1}^r) = M\ LP_{\theta_l}(s_t)\ \ (s_t\text{: decoder state at time } t)$$

- Context processor
  - Attend context: $\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_i \exp(e_{t,i})}$, $c_t^u = \sum_{i=1}^{l_u} \alpha_{t,i} h_i^u$ ($e_{t,i}$: similarity between s_t and h_i^u)
  - The generation prob is defined by:

$$P(w_t^r | U, w_{1:t-1}^r) = p_{gen} P_{vocab}(w_t^r | U, w_{1:t-1}^r) + (1 - p_{gen}) \sum_{i: w_i^u = w_i^r} \alpha_{t,i}$$

- Context encoder
  - Transform context tokens into a sequence of hidden vectors:

$$h_1^u, \dots, h_i^u, \dots, h_{l_u}^u = GRU_{\theta_c}(e_1^u, \dots, e_i^u, \dots, e_{l_u}^u)\ (e_i^u\text{: } w\ ordem\ \ bedding)$$

Pre-trained with the Plain Text

- Knowledge encoder
  - Represent d_i as a sequence of hidden vectors:

$$h_{i,1}^{d}, \ldots, h_{i,j}^{d}, \ldots, h_{i,l_d}^{d} = BiGRU_{\theta_k}(e_{i,1}^{d}, \ldots, e_{i,j}^{d}, \ldots, e_{i,l_d}^{d})$$

Learned on the KG Dialogues

- Knowledge Processor
  - Attend knowledge:

$$\beta^s_{t,i} = \frac{\exp\left(g_{\theta_{s'}}\left(s_t, \hat{h}^d_i\right)\right)}{Z_s}; \beta^w_{t,i,j} = \frac{\exp\left(g_{\theta_{s'}}\left(s_t, h^d_{i,j}\right)\right)}{Z_w}$$ (Z_s, Z_w:normalization factors; $g_{\theta_{s'}}$:similarity function parametered by $\theta_{s'}$)

  - Knowledge vector:

$$c^d_t = \sum_{i=1}^{m} \beta^s_{t,i} \hat{h}^d_i$$

  - The generation prob is formulated as:

$$P(w^r_t | D, w^r_{1:t-1}) = p'_{gen} P_{vocab}(w^r_t | D, w^r_{1:t-1}) + (1 - p'_{gen}) \sum_{i,j: w^d_{i,j} = w^r_t} \beta_{t,i,j}$$

- Decoding Manager
  - The prob to predict word w^r_t can be formulated as:

$$P(w^r_t | U, D, w^r_{1:t-1}) = \pi_t \cdot \begin{bmatrix} P(w^r_t | w^r_{1:t-1}); \\ P(w^r_t | U, w^r_{1:t-1}); \\ P(w^r_t | D, w^r_{1:t-1}) \end{bmatrix}$$ (π_t: distribution of the three componets, one-hot vector)

$$\pi_t = gumbel\_softmax(f_\pi(s_{t-1}), \tau) \in \mathbb{R}^{1\times 3}$$ (f_π: MLP; s_{t-1}: decoder state at time t-1;τ: temperature)

Experiments

- Datasets
  - Wizard of Wikipedia (facebook): Test Seen vs Test Unseen
  - CMU_DoG (cmu)
- Evaluation Metrics
  - PPL, F1, BLEU, BOW Embedding
  - Human evaluation
- Baselines
  - Transformer Memory Network (TMN), ICLR 2019
  - Incremental Transformer with Deliberation Decoder (ITDD), ACL 2019

Evaluation Results

- Our model only needs 1/8 training data to achieve the state-of-the-art performance
- The model outperforms the baseline models on out-of-domain knowledge
- When the training set shrinks, the performance gap on Test Seen and Test Unseen becomes marginal
  - Better capability for generalization

Models \ Metrics	PPL	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Average	Extrema	Greedy
TMN (Dinan et al., 2019)	66.5	15.9	0.184	0.073	0.033	0.017	0.844	0.427	0.658
ITDD (Li et al., 2019)	17.8	16.2	0.158	0.071	0.040	0.025	0.841	0.425	0.654
FULL DATA	23.0	18.0	0.218	0.115	0.075	0.055	0.835	0.434	0.658
1/2 DATA	25.3	17.5	0.217	0.113	0.073	0.053	0.833	0.431	0.657
1/4 DATA	29.2	16.9	0.212	0.105	0.064	0.044	0.833	0.429	0.658
1/8 DATA	33.5	16.3	0.206	0.098	0.059	0.039	0.832	0.425	0.658
1/16 DATA	38.6	15.7	0.197	0.091	0.052	0.033	0.834	0.428	0.655

Table 1: Evaluation results on Test Seen of Wizard.

Models \ Metrics	PPL	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Average	Extrema	Greedy
TMN (Dinan et al., 2019)	103.6	14.3	0.168	0.057	0.022	0.009	0.839	0.408	0.645
ITDD (Li et al., 2019)	44.8	11.4	0.134	0.047	0.021	0.011	0.826	0.364	0.624
FULL DATA	25.6	16.5	0.207	0.101	0.062	0.043	0.828	0.422	0.628
1/2 DATA	27.7	16.7	0.208	0.103	0.064	0.045	0.827	0.421	0.647
1/4 DATA	32.4	16.2	0.205	0.098	0.060	0.041	0.828	0.423	0.650
1/8 DATA	35.8	16.0	0.201	0.093	0.054	0.035	0.831	0.419	0.653
1/16 DATA	41.0	15.3	0.191	0.087	0.050	0.032	0.832	0.424	0.652

Table 2: Evaluation results on Test Unseen of Wizard.

Evaluation Results

- Evaluation results on CMU_DoG
  - Similar findings: 1/8 data size with better results than SOTA

Models \ Metrics	PPL	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Average	Extrema	Greedy
TMN (Dinan et al., 2019)	75.2	9.9	0.115	0.040	0.016	0.007	0.789	0.399	0.615
ITDD (Li et al., 2019)	26.0	10.4	0.095	0.036	0.017	0.009	0.748	0.390	0.587
FULL DATA	54.4	10.7	0.150	0.057	0.025	0.012	0.809	0.413	0.633
1/2 DATA	57.0	10.4	0.142	0.052	0.022	0.010	0.808	0.414	0.635
1/4 DATA	61.7	10.5	0.131	0.046	0.019	0.009	0.781	0.402	0.613
1/8 DATA	67.6	10.2	0.121	0.044	0.019	0.009	0.787	0.407	0.622

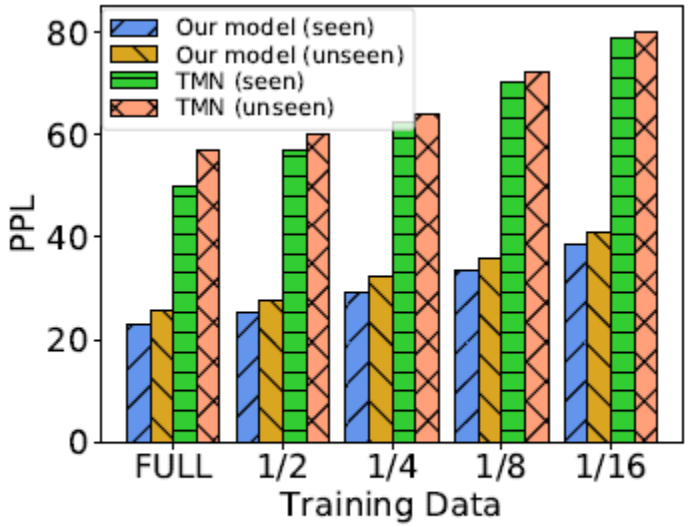
Table 3: Evaluation results on CMU_DoG.

Comparison of Parameter Fine-Tuning and Parameter Fixing

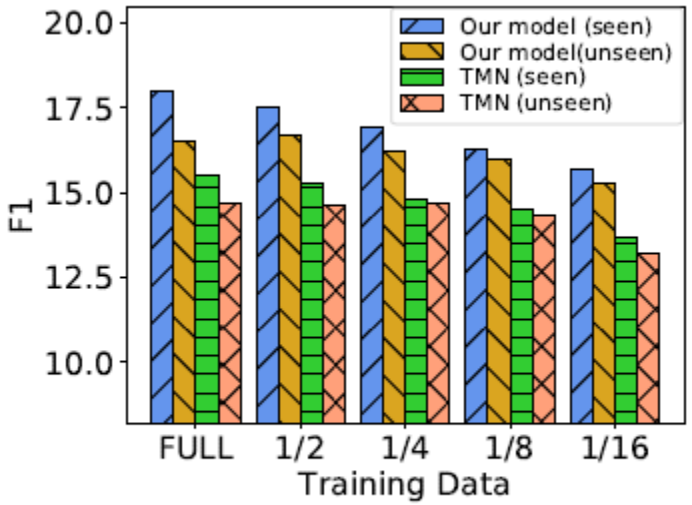
- Fine-tuning can further improve the model on both in-domain and out-of-domain knowledge when there are enough training data.

Comparison with Pre-trained TMN

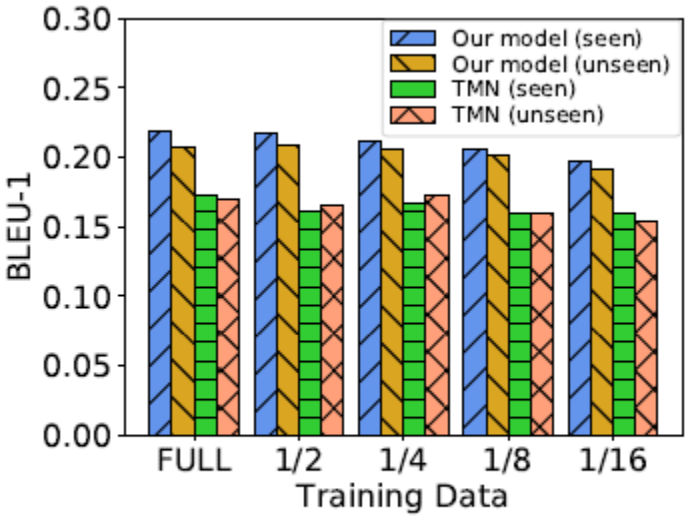
- Entangling (TMN) vs. disentangling (Ours)
  - Disentangling is important to leverage ungrounded dialogues and plain text for low-resource knowledge-grounded dialogue generation



(a) PPL



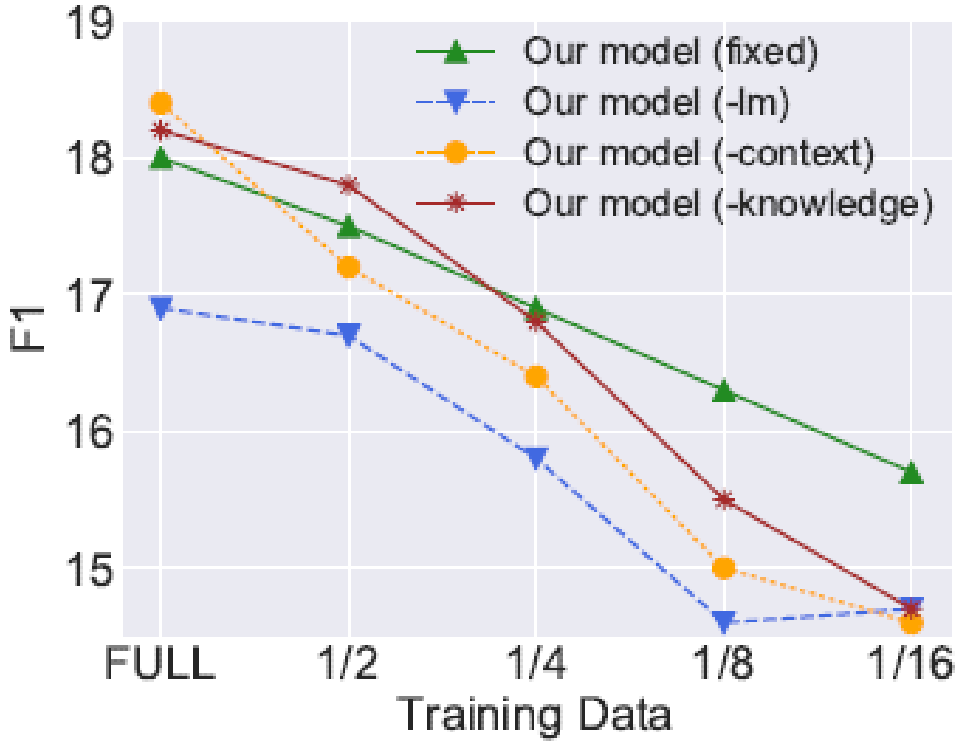
(b) F1

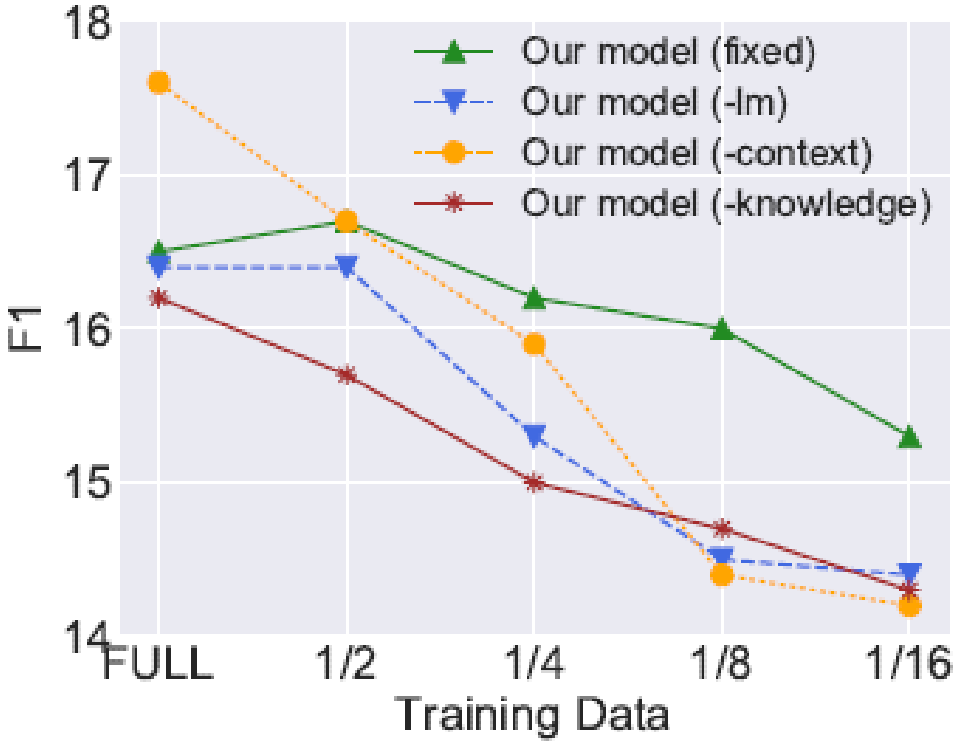


(c) BLEU-1

Results of Pre-training Ablation

- Removing any component from pre-training causes performance drop when training data is small.





Case Study

[Knowledge]	✧ the lineup , when first signed to geffen records in 1986 , consisted of vocalist axl rose , lead guitarist slash , rhythm guitarist izzy stradlin , bassist duff mckagan , and drummer steven adler . ✧ guns n ' roses has released six studio albums , accumulating sales of more than 100 million records worldwide , including 45 million in the united states , making them the 41st best-selling artist of all time . ✧ guns n ' roses ' debut album , `` appetite for destruction '' ( 1987 ) , reached number one on the `` billboard '' 200 a year after its release , on the strength of `` sweet child o ' mine '' , the band 's only single to reach number one on the `` billboard '' hot 100 . ✧ guns n ' roses , often abbreviated as gnr , is an american hard rock band from los angeles , california , formed in 1985 . ✧ the album has sold approximately 30 million copies worldwide , including 18 million units in the united states , making it the best-selling debut album of all time in the us , as well as the eleventh best-selling album in the united states . ✧ the success of the debut was followed by the eight-song album `` g n ' r lies '' ( 1988 ) which reached number two on the `` billboard '' 200 . ✧ the twin albums `` use your illusion i '' and `` use your illusion ii '' (1991) debuted at number two and number one on the `` billboard '' 200 respectively and have sold a combined 35 million copies worldwide , including 14 million units in the united states .
[Context]	have you heard of the band guns n ' roses ? they are my favorite band ever .
[Groundtruth]	sure the band was formed inn 1985 in los angeles , california . what is your favorite album ?
[TMN]	they are a great band .
[ITDD]	i have n't , but i do know that they are the most popular band in the us .
Our Model	me too ! they 're an american rock band from los angeles and has released six studio albums .

Generated by Language model Generated by Context Processor Generated by Knowledge Processor

Table 6: A case from Test Unseen of Wizard.

Summary

- A knowledge-grounded dialogue generation model under a low-resource setting
- A new model for decoupling knowledge and dialogues
  - Most parameters (~95%) can be estimated from ungrounded dialogues and documents
  - Decomposing the response decoder into independent components
- Outperform SOTA methods with only 1/8 data

中国人民大学 RENMIN UNIVERSITY OF CHINA 高瓴人工智能学院 Gaoling School of Artificial Intelligence

Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao and **Rui Yan**

AAAI 2021

Motivations

- Existing studies focus on building a context-response matching model with various neural architectures or pre-trained language models (PLMs) and typically learning with a single response prediction task
  - Still challenging to effectively learn task-related knowledge during the training process, especially when the size of training corpora is limited.
  - Overlook many potential training signals contained in dialogue data.
- Response retrieved systems supervised by the conventional way still faces some from existing dialogue critical challenges
  - Including incoherence and inconsistency

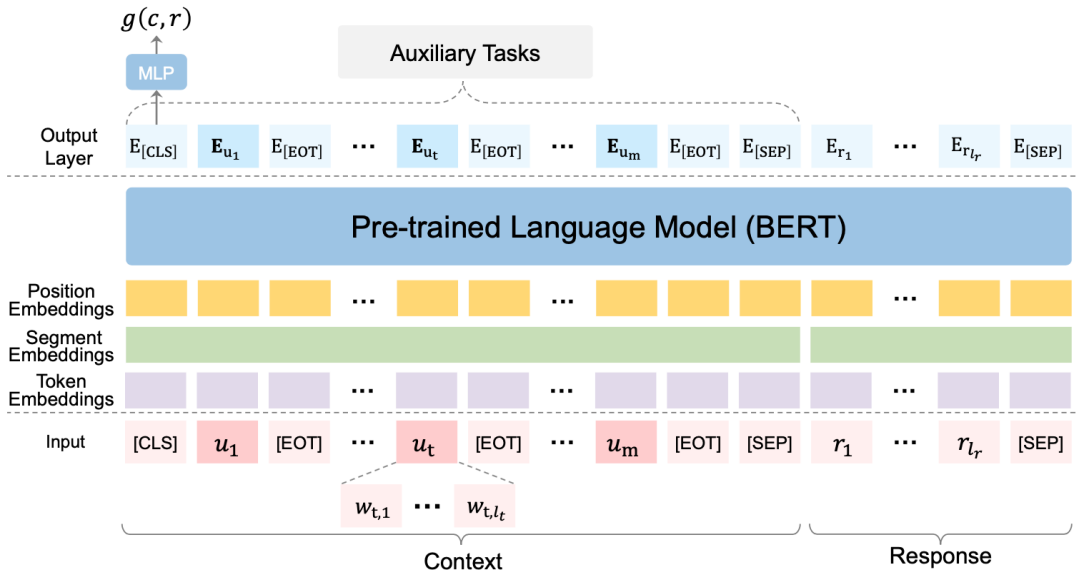
Methods

- How to learn an effective context-response matching model with limited corpora?

- Conventional Approach:

$$\mathcal{L}_{\text{crm}} = -y \log(g(c,r)) - (1-y)\log(1-g(c,r))$$

- Self-supervised learning:
  - Constructing various training signals with multi-turn dialogue.
  - Jointly training with response matching task.



Better task-related representation
Better generalization ability

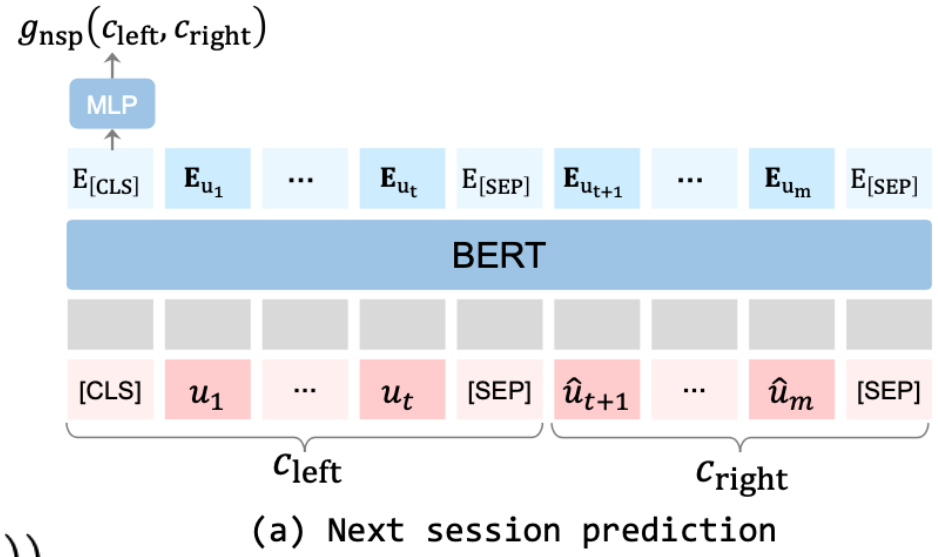
Self-Supervised Tasks

a) Next Session Prediction (NSP)

- Predict whether two pieces of dialogue session are consecutive and relevant

$$c_{left} = \{u_1, \dots u_t\} \qquad c_{right} = \{u_t, \dots u_m\}$$

$$\mathcal{L}_{\text{nsp}} = -y_{\text{nsp}} \log(g_{\text{nsp}}(c_{\text{left}}, c_{\text{right}})) - (1 - y_{\text{nsp}}) \log(1 - g_{\text{nsp}}(c_{\text{left}}, c_{\text{right}}))$$



(a) Next session prediction

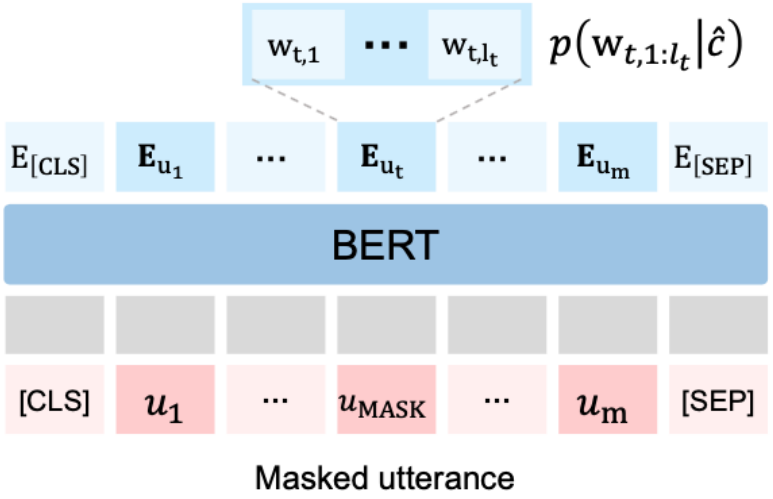
Self-Supervised Tasks

b) Utterance restoration

$$E'_{t,j} = \text{GLEU}(W_{ur}E_{t,j} + b_{ur})$$

$$p(w_{t,j}|\hat{c}) = \text{softmax}\left(W'_{ur}E'_{t,j} + b'_{ur}\right)$$

$$\mathcal{L}_{\text{ur}} = -\frac{1}{l_t}\sum_{j=1}^{l_t}\log p(w_{t,j}|\hat{c})$$



(b) Utterance restoration

Self-Supervised Tasks

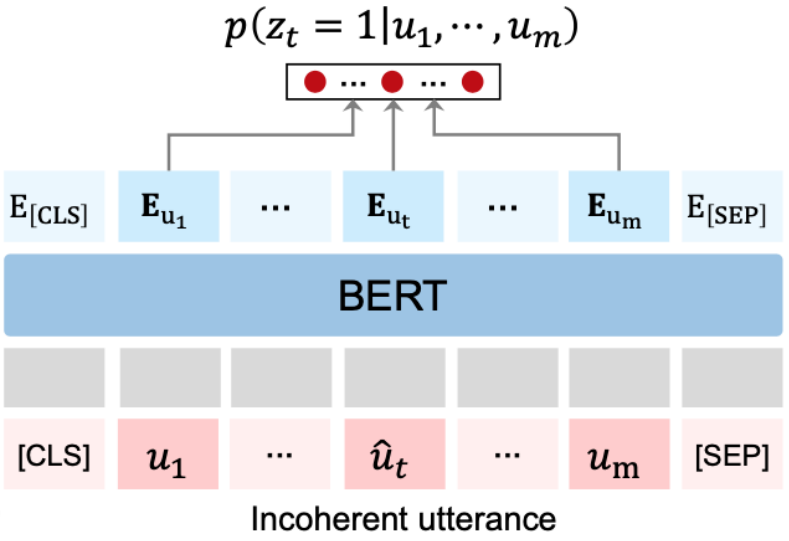
c) Incoherence Detection (ID)

- Recognize the incoherent utterance (randomly replaced) within a dialogue session

$$U_t = [\frac{1}{l_t}\sum_{j=1}^{l_t} E_{t,j}; \max_{1\le j\le l_t} E_{t,j}]$$

$$p(z_t = 1|u_1, \ldots, u_m) = \text{softmax}(W_{\text{id}} U_t + b_{\text{id}})$$

$$\mathcal{L}_{\text{id}} = -\sum_{t=1}^{m} z_t \log p(z_t = 1|u_1, \ldots, u_m)$$



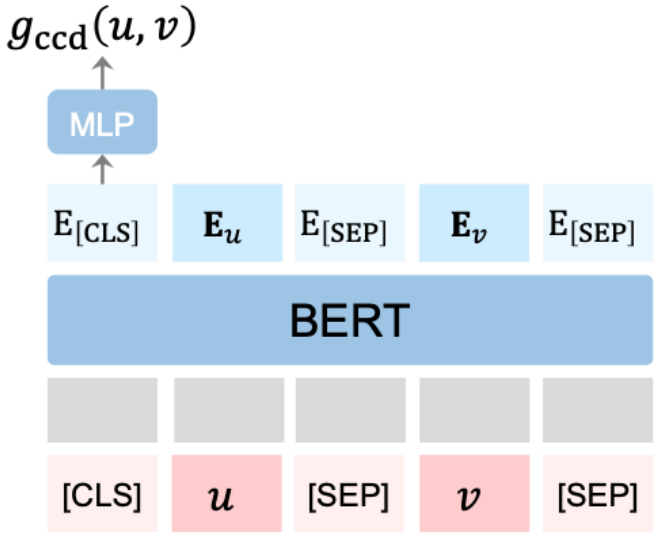
(c) Incoherence detection

Self-Supervised Tasks

d) Consistency Discrimination (CD)

- Measure the consistency among two utterances
- u and v are from the same interlocutor in the same dialogue session, $\tilde{v}$ is from other session.

$$\mathcal{L}_{\text{cd}} = \max\{0, \Delta - g_{\text{cd}}(u, v) + g_{\text{cd}}(u, \tilde{v})\}$$



(d) Consistency discrimination

Learning Objective

- Multi-task learning:
  - $\mathcal{L}_{final} = \mathcal{L}_{crm} + \alpha\mathcal{L}_{self}$
  - $\mathcal{L}_{self} = \mathcal{L}_{nsp} + \mathcal{L}_{ur} + \mathcal{L}_{id} + \mathcal{L}_{cd}$

(α is the trade-off between the objective of the main task and. those of the auxiliary tasks)

Datasets

- Ubuntu Dialogue Corpus
  - Multi-turn English dialogues about technical support and is collected from chat logs of the Ubuntu forum
- E-commerce Dialogue Corpus
  - Real-world multi-turn dialogues between customers and customer service staff on Taobao, the largest e-commerce platform in China

dataset	Ubuntu Corpus			E-Commerce Corpus		
	Train	Dev	Test	Train	Dev	Test
# context-response pairs	1M	500k	500k	1M	10K	10K
# candidates per context	2	10	10	2	2	10
# positive candidates per context	1	1	1	1	1	1
Avg. # turns per dialogue	10.13	10.11	10.11	5.51	5.48	5.64
Avg. # words per dialogue	136.91	136.49	136.92	46.57	46.12	50.76

Experiment Results

Models \ Metrics		Ubuntu Corpus				E-commerce Corpus		
		$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
Non-PLM-based Models	DualLSTM (Lowe et al. 2015)	0.901	0.638	0.784	0.949	0.365	0.536	0.828
	Multi-View (Zhou et al. 2016)	0.908	0.662	0.801	0.951	0.421	0.601	0.861
	SMN (Wu et al. 2017)	0.926	0.726	0.847	0.961	0.453	0.654	0.886
	DUA (Zhang et al. 2018)	-	0.752	0.868	0.962	0.501	0.700	0.921
	DAM (Zhou et al. 2018)	0.938	0.767	0.874	0.969	0.526	0.727	0.933
	MRFN (Tao et al. 2019a)	0.945	0.786	0.886	0.976	-	-	-
	IMN (Gu, Ling, and Liu 2019)	0.946	0.794	0.889	0.974	0.621	0.797	0.964
	ESIM (Chen and Wang 2019)	0.950	0.796	0.874	0.975	0.570	0.767	0.948
	IoI (Tao et al. 2019b)	0.947	0.796	0.894	0.974	0.563	0.768	0.950
	MSN (Yuan et al. 2019)	-	0.800	0.899	0.978	0.606	0.770	0.937
PLM-based Models	BERT (Whang et al. 2020)	0.952	0.814	0.902	0.977	0.631	0.826	0.964
	SA-BERT (Gu et al. 2020)	0.965	0.855	0.928	0.983	0.704	0.879	0.985
	BERT-VFT (Whang et al. 2020)	-	0.855	0.928	0.985	-	-	-
	BERT-VFT (Ours)	0.969	0.867	0.939	0.987	0.717	0.884	0.986
	BERT-SL	**0.975***	**0.884***	**0.946***	**0.990***	**0.776***	**0.919***	0.991
	BERT-SL w/o. NSP	0.973	0.879	0.944	0.989	0.760	0.914	0.988
	BERT-SL w/o. UR	0.974	0.881	0.945	0.990	0.763	0.916	0.991
	BERT-SL w/o. ID	0.972	0.877	0.942	0.989	0.755	0.911	0.987
	BERT-SL w/o. CD	0.973	0.880	0.945	0.989	0.742	0.897	0.986

Experiment Results

- Self-supervised learning for non-pretrained models

Metrics / Models	Ubuntu Corpus			
	$R_2@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
DualLSTM (Lowe et al. 2015)	0.901	0.638	0.784	0.949
DualLSTM-SL	0.925*	0.724*	0.858*	0.969*
ESIM (Chen and Wang 2019)	0.950	0.796	0.874	0.975
ESIM-SL	0.963*	0.822*	0.909*	0.980*
BERT (Devlin et al. 2019)	0.954	0.817	0.904	0.977

One-ninth of the parameter of BERT

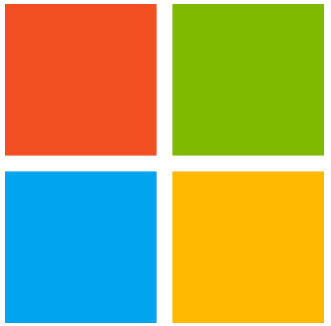
Experiment Results

- Human Evaluation

Models \ Metrics	Relevance	Coherence	Consistency	Fleiss' kappa
MSN [37]	1.55	1.45	1.55	0.675
BERT [31]	1.58	1.44	1.58	0.714
BERT-VFT [31]	1.64	1.51	1.61	0.681
BERT-SL (Our)	1.65	1.63	1.66	0.635

Summary

- We propose learning a context-response matching model with multiple auxiliary self-supervised tasks designed for the dialogue data
- Jointly trained with these self-supervised tasks, the matching model can effectively learn task-related knowledge contained in dialogue data, achieve a better local optimum and produce better features for response selection
- Experiment results on two benchmarks indicate that the proposed auxiliary self-supervised tasks bring significant improvement for various matching architectures on multi-turn response selection in retrieval-based dialogues
  - New state-of-the-art results on both datasets



A Pre-training Strategy for Zero-Resource Response Selection in Knowledge-Grounded Conversations

Chongyang Tao, Changyu Chen, Jiazhan Feng, Ji-Rong Wen and **Rui Yan**

ACL 2021

Knowledge-Grounded Response Selection

- Given a conversation context and a set of knowledge entries
  - ① Select proper knowledge
  - ② Distinguish the most appropirate response from a candidate pool

Background Knowledge	
Name	The inception
Year	2009
Director	Christopher Nolan
Genre	Scientific
Cast	Leonardo DiCaprio as Dom Cobb, a professional thief who specializes in conning secrets from his victims by infiltrating their dreams. Tom Hardy as Eames, a sharp-tongued associate of Cobb. ...
Introd--uction	Dominick Cobb and Arthur are extractors, who perform corporate espionage using an experimental military technology to infiltrate the subconscious of their targets and extract valuable information through a shared dream world. Their latest target, Japanese businessman Saito, reveals that he arranged the mission himself to test Cobb for a seemingly impossible job: planting an idea in a person's subconscious, or inception.
Rating	Rotten Tomatoes: 86% and average: 8.1/10; IMDB: 8.8/10
Conversation	
User 2:	Hi how are you today?
User 1:	I am good. How are you?
User 2:	Pretty good. Have you seen the inception?
User 1:	No, I have not but have heard of it. What is it about?
User 2:	It's about extractors that perform experiments using military technology on people to retrieve info about their targets.
User 1:	Sounds interesting. Do you know which actors are in it?
User 2:	I haven't watched it either or seen a preview. But it's scifi so it might be good. Ugh Leonardo DiCaprio is the main character.
User 2:	He plays as Don Cobb.
User 1:	I'm not a big scifi fan but there are a few movies I still enjoy in that genre.
User 1:	Is it a long movie?
User 2:	Doesn't say how long it is.
User 2:	The Rotten Tomatoes score is 86%.

Challenges

- It is non-trivial to collect large-scale dialogues that are naturally grounded on the background documents.

- *Can we learn a knowledge-grounded response selection model without any knowledge-grounded dialogue data?*

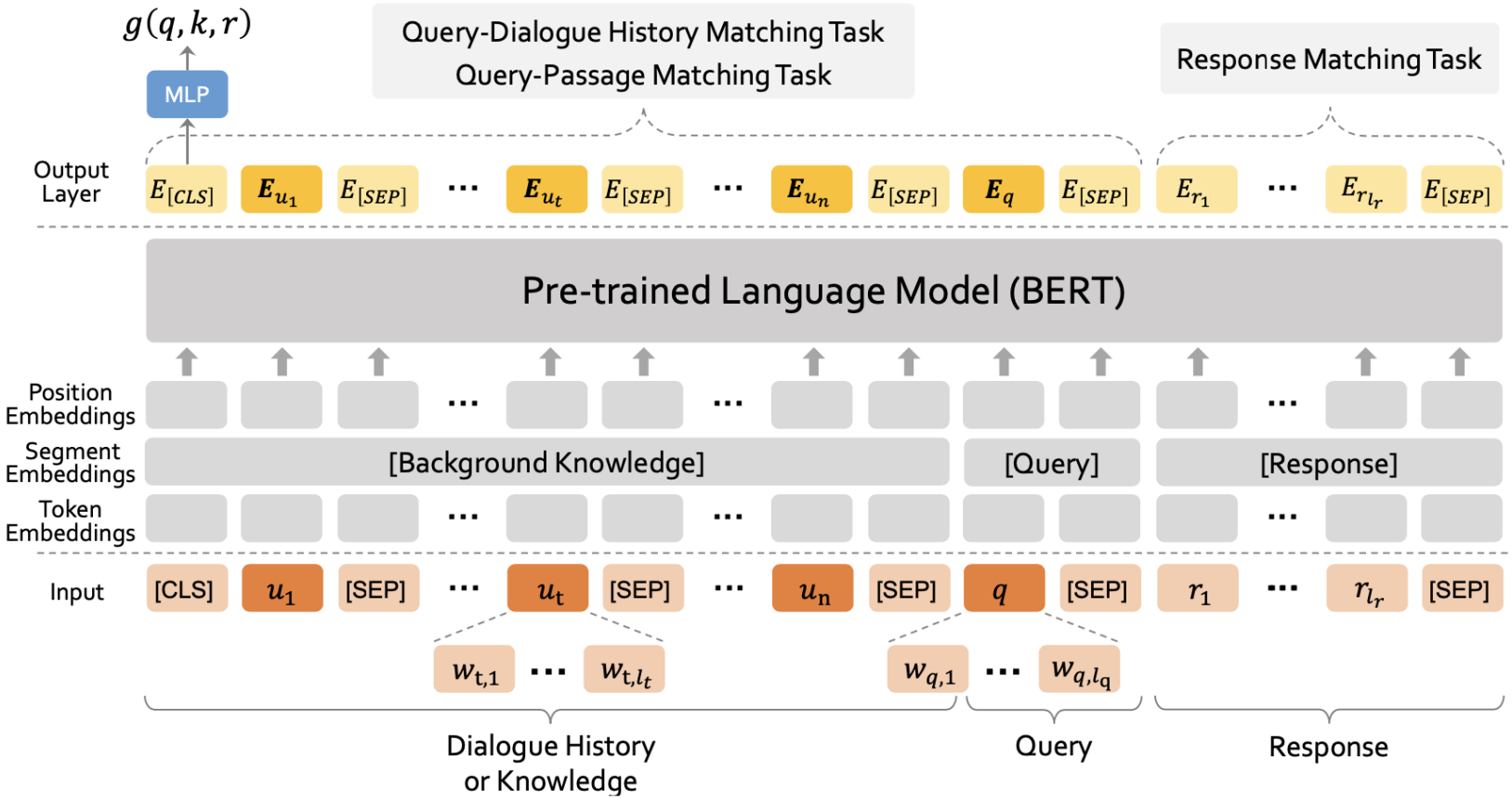
Motivations

dialogue history, query

- Since <knowledge, context, response > triples are hard to collect, but the following data is abundant
  - unstructured knowledge (e.g., web pages or wiki articles) <knowledge>
  - passage search datasets (e.g., ad-hoc retrieval tasks) <query, knowledge>
  - multi-turn dialogues (e.g., Reddit) < query, dialogue history, knowledge >
- The *background knowledge* and *dialogue history* are *symmetric* in terms of the information they convey, and we assume that the dialogue history can be regarded as another format of background knowledge for response prediction.

Model Overview

- Decomposing the training of the grounded response selection task into several sub-tasks, and joint learning all those tasks in a unified model.



Pre-training Strategies

- Task1 - Query-Passage Matching
  - Predict whether a query and the the passage and relevant.

$$S^{qp} = \{\texttt{[CLS]}, w_1^p, \ldots, w_{n_p}^p, \texttt{[SEP]}, w_1^q, \ldots, w_{n_q}^q\}$$

  - Loss function

$$\mathcal{L}_{\mathrm{p}}(q, p^+, p_1^-, \ldots, p_{n_p}^-) = -\log\left(\frac{e^{g(q,p^+)}}{e^{g(q,p^+)} + \sum_{j=1}^{\delta_p} e^{g(q,p_j^-)}}\right)$$

  - where p^+ stands for the positive passage for q, p_j^- is the j-th negative passage and δ_p is the number of negative passage.

Pre-training Strategies

- Task2 - Query-Dialogue History Matching
  - Predict whether a query and a dialogue are consecutive and relevant.

$$S^{qh} = \{\texttt{[CLS]}, w_1^h, \ldots, w_{n_h}^h, \texttt{[SEP]}, w_1^q, \ldots, w_{n_q}^q\}$$

  - Loss function

$$\mathcal{L}_{\mathrm{h}}(q, h^+, h_1^-, \ldots, h_{n_h}^-) = -\log\left(\frac{e^{g(q,h^+)}}{e^{g(q,h^+)} + \sum_{j=1}^{\delta_h} e^{g(q,h_j^-)}}\right)$$

  - where h^+ is the true dialogue history for q, h_j^- is the j-th negative dialogue history randomly sampled from the training set and δ_h is the number of negative passage.

Pre-training Strategies

- Task3 - Multi-turn Response Matching
  - Predict whether a response candidate is appropriate for a given query and a concatenated dialogue history sequence

$$S^{hqr} = \{\texttt{[CLS]}, w_1^h, \ldots, w_{n_h}^h, \texttt{[SEP]}, w_1^q, \ldots, w_{n_q}^q, \texttt{[SEP]}, w_1^r, \ldots, w_{n_r}^r\}$$

  - Loss function

$$\mathcal{L}_{\mathbf{r}}(h, q, r^+, r_1^-, \ldots, r_{\delta_r}^-) = -\log\left(\frac{e^{g(h,q,r^+)}}{e^{g(h,q,r^+)} + \sum_{i=j}^{n_r} e^{g(h,q,r_j^-)}}\right)$$

  - where r^+ is the true response for a given q and h, r_-^j is the j-th negative response candidate and δ_r is the number of negative response candidate.

Inference

- First rank $\{p_i\}_{i=1}^{n_k}$ according to $g(q, k_i)$ and then select top m knowledge entries $\{p_1, \ldots, p_m\}$ for the subsequent response matching process.
- Present two strategies to compute the final matching score g(k, h, q, r)
  - Cat:

$$g(k, h, q, r) = g(p_1 \oplus \ldots \oplus p_m \oplus c, q, r)$$

  - Sep:

$$g(k, h, q, r) = g(h, q, r) + \max_{i \in (0,m)} g(p_i, q, r)$$

Experimental Setup

- Training set
  - MS MARCO passage ranking dataset (Nguyen et al., 2016)
    - 500k pairs of query and relevant passage.
    - Another 400M passages are used as the sampling pool.
  - Reddit dialogue dataset (Dziri et al., 2018)
    - Randomly sampling 2.28M/20K dialogues as the training/validation set.
    - On average, each dialogue contains 4.3 utterances, and the average length of the utterances is 42.5.

Experimental Setup

- Testing set
  - We tested our proposed method on the Wizard- of-Wikipedia (WoW) (Dinan et al., 2019) and CMU DoG (Zhou et al., 2018a).

Statistics	Wizard of Wikipedia		CMU_DoG
	Test Seen	Test Unseen	Test
Avg. # turns	9.0	9.1	12.4
Avg, # words per turn	16.4	16.1	18.1
Avg. # knowledge entries	60.8	61.0	31.8
Avg. # words per knowledge	36.9	37.0	27.0

- Evaluation metrics
  - $R_n@k$

Baselines

- IR Baseline (Dinan et al., 2019)
- BoW MemNet (Dinan et al., 2019)
  - Learn the knowledge selection and response matching jointly
- Transformer MemNet (Dinan et al., 2019)
  - A pretrained Transformer
  - Learn the knowledge selection and response matching jointly
- Two-stage Transformer (Dinan et al., 2019)
- DIM (Gu et al., 2019)
- FIRE (Gu et al., 2020)

Evaluation Results

- The performance of knowledge selection

Supervised Methods

Models	Test Seen			Test Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
IR Baseline	17.8	-	-	14.2	-	-
BoW MemNet	71.3	-	-	33.1	-	-
Two-stage Transformer	84.2	-	-	63.1	-	-
Transformer MemNet	87.4	-	-	69.8	-	-
DIM (Gu et al., 2019)	83.1	91.1	95.7	60.3	77.8	92.3
FIRE (Gu et al., 2020b)	88.3	95.3	97.7	68.3	84.5	95.1
$\text{PTKGC}_{\text{cat}}$	85.7	94.6	98.2	65.5	82.0	94.7
$\text{PTKGC}_{\text{sep}}$	89.5	96.7	98.9	69.6	85.8	96.3

Evaluation results on the test set of WoW.

Models	R@1	R@2	R@5
Starspace (Wu et al., 2018)	50.7	64.5	80.3
BoW MemNet (Zhang et al., 2018)	51.6	65.8	81.4
KV Profile Memory (Zhang et al., 2018)	56.1	69.9	82.4
Transformer MemNet (Mazaré et al., 2018)	60.3	74.4	87.4
DGMN (Zhao et al., 2019)	65.6	78.3	91.2
DIM (Gu et al., 2019)	78.7	89.0	97.1
FIRE (Gu et al., 2020b)	81.8	90.8	97.4
$\text{PTKGC}_{\text{cat}}$	61.6	73.5	86.1
$\text{PTKGC}_{\text{sep}}$	66.1	77.8	88.7

Evaluation results on the test set of CMU_DoG.

Evaluation Results

- The performance of knowledge selection

Models	Wizard Seen			Wizard Unseen		
	R@1	R@2	R@5	R@1	R@2	R@5
Random	2.7	-	-	2.3	-	-
IR Baseline	5.8	-	-	7.6	-	-
BoW MemNet	23.0	-	-	8.9	-	-
Transformer	22.5	-	-	12.2	-	-
Transformer (w/ pretrain)	25.5	-	-	22.9	-	-
Our Model	22.0	31.2	48.8	23.1	32.1	50.7

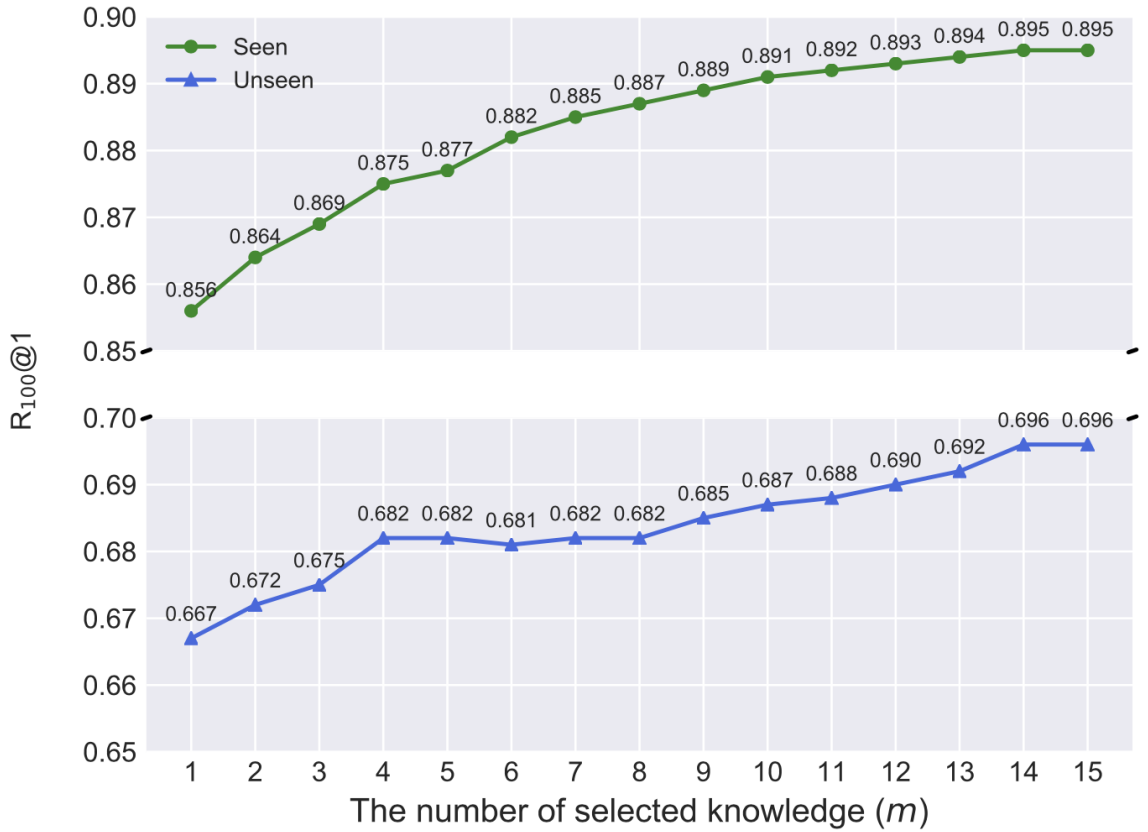
Evaluation Results

- Ablation studies

Models	Wizard of Wikipedia						CMU_DoG		
	Test Seen			Test Unseen					
	R@1	R@2	R@5	R@1	R@2	R@5	R@1	R@2	R@5
$\text{PTKGC}_{\text{sep}}$	89.5	96.7	98.9	69.6	85.8	96.3	66.1	77.8	88.7
$\text{PTKGC}_{\text{sep}}$ (q)	70.6	79.7	86.8	55.9	70.8	83.4	47.3	58.8	75.0
$\text{PTKGC}_{\text{sep}}$ (q+h)	84.9	93.9	97.8	64.9	81.7	94.3	59.5	72.3	86.1
$\text{PTKGC}_{\text{sep}}$ (q+k)	89.5	96.4	98.6	67.0	84.0	96.0	62.7	73.8	84.8
$\text{PTKGC}_{\text{sep,m=1}}$	85.6	94.4	97.9	66.7	82.8	94.3	60.4	72.5	86.0
$\text{PTKGC}_{\text{sep,m=1}}$ - $\mathcal{L}_{\text{p}}$	84.7	93.5	97.5	63.4	80.5	94.0	58.7	70.8	85.6
$\text{PTKGC}_{\text{sep,m=1}}$ - $\mathcal{L}_{\text{h}}$	84.9	93.7	97.6	65.5	81.7	94.1	59.4	71.4	85.3

Discussion

- The performance of response selection across different number of selected knowledge.



Summary

- Exploration of response matching in knowledge-grounded conversations under a zero- resource setting
- Proposal of decomposing the training of the knowledge-grounded response selection into three tasks and joint train all tasks in a unified pre-trained language model.
- Exiperimental results on two benchmarks demonstrate the effectiveness of our method.

Reasoning in Dialog:

Improving Response Generation by Context Reading Comprehension

Xiuying Chen, Zhi Cui, Jiayi Zhang, Chen Wei, Jianwei Cui, Bin Wang, Dongyan Zhao and **Rui Yan**

AAAI 2021

Motivation

- In multi-turn dialog, utterances do not always take the full form of sentences (Carbonell 1983)
  - Understanding the dialog context is difficult
- Essential to fully grasp the dialog context to generate a reasonable response
- Goal: to improve the response generation performance
  - Examining the model's ability to answer a reading comprehension question
  - The question is focused on the omitted information in the dialog

Motivation

	Example 1	Example 2	Example 3
A_1	求帮忙取名字姓程，俩男娃 Please help me decide how to name my two kids whose last name is Cheng	我最喜欢的歌手是MJ My favorite singer is MJ	那么我们即使不死,也在天堂 Then we are in heaven even if we don't die
B_1	程饭和程菜 Cheng fan and Cheng cai	你最喜欢他的什么歌呢? What's your favorite song?	这话哪抄的 Where did you copy that
A_2	哈哈哈哈哈 LOL	Thriller Thriller	三毛 Sanmao
B_2	请务必接受我的建议 Please accept my advice	我没听过呢，有这首歌的mv吗 I haven't heard of it. Is there an MV of this song?	想起以前豆瓣有个帅哥叫东门 Remember that there was a handsome man named Dongmen in Douban
A_3	咱俩一起生我就接受（取名程饭和程菜） I'll accept that (name as Cheng fan and Cheng cai) if they are our children	有(这首歌的MV)，我发给你看 Yes (I have the MV), I'll send it to you	那我（豆瓣）叫个南亭算了 Then my name (in Douban) will be Nanting.
Question	如果一起生娃那孩子叫什么 If the children are ours, how to call them?	准备发什么? What is going to be sent?	南亭是什么的ID? Nanting is ID of what
Answer	程饭和程菜 Cheng fan and Cheng cai	发thriller的的mv Thriller MV	豆瓣 Douban
B_3	我觉得这名字很好听啊 I think it's a nice name	好啊，我一直想看他的MV呢 Good, I've always wanted to see his MV	豆瓣就差你这个ID了 Douban is waiting for your ID
Reasoning Type	Paraphrasing (49.0%)	Lexical match (28.5%)	Pragmatics (22.5%)

Table 1: Examples from the dataset. Questions are concentrated on the omitted information of A_3 (which is shown in brackets), and reasoning type is the type of ability that is needed to answer the question.

Contribution

- A multi-task learning framework:
  - Jointly answers reading comprehension questions and generates a proper response in multi-turn dialog scenario
- New method:
  - Transformer architecture with a memory updater
  - Selectively store and update history dialog information
- New data
  - A large scale dialog reading comprehension dataset
- Experimental results on this dataset demonstrate the effectiveness

Overview

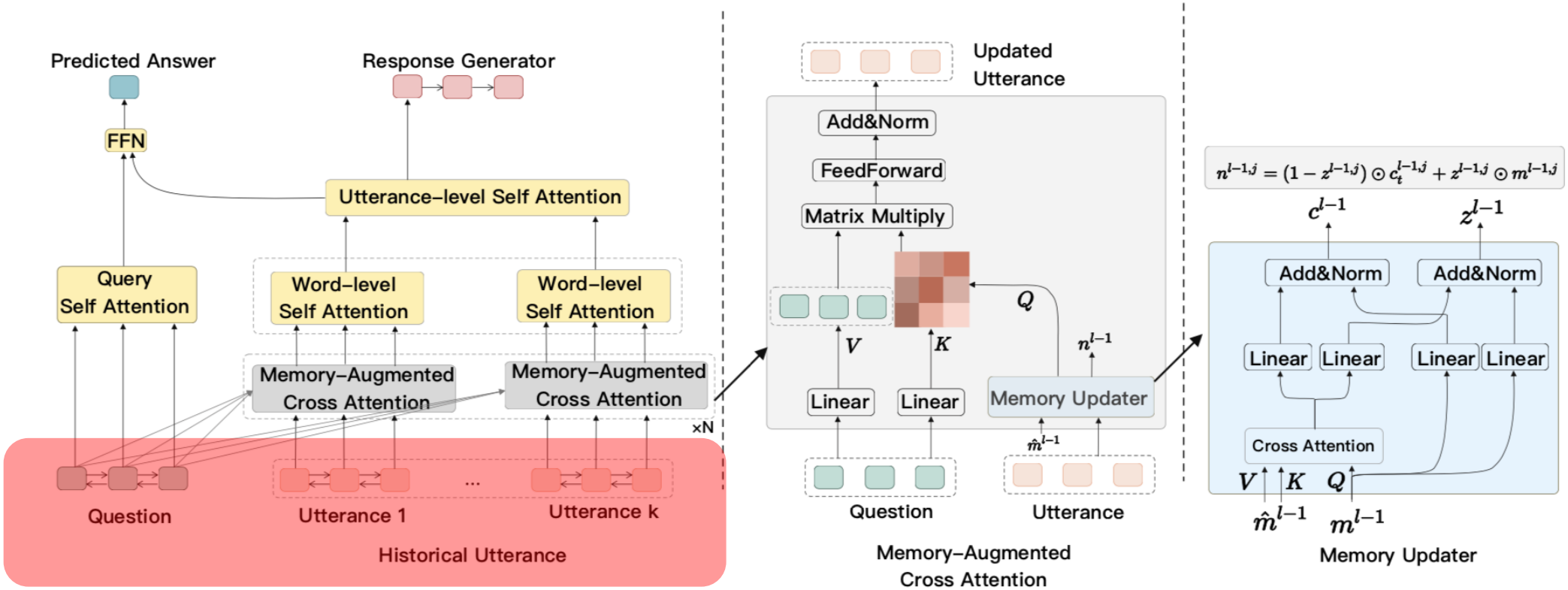
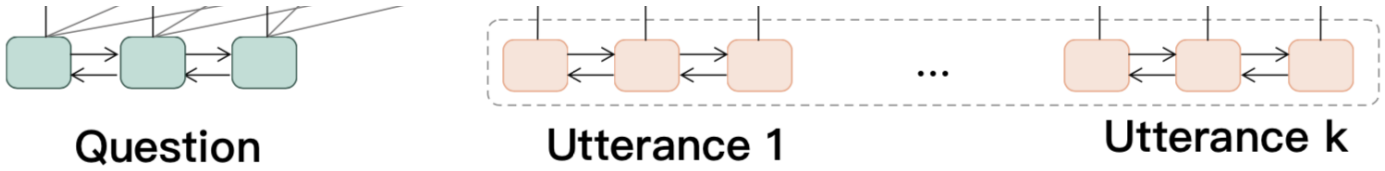


Figure 1: Overview of MRG. We divide our model into three parts: (1) Cross-hierarchical Encoder (which consists of memory-augment cross attention and two hierarchical self attentions); (2) Answer Selecter; (3) Response Generator.

Cross-Hierarchical Encoder



$$h_i^{x,j} = \text{Bi-RNN}_x\left(e(x_i^j), h_{i-1}^{x,j}\right),$$
$$h_i^q = \text{Bi-RNN}_y\left(e(q_i), h_{i-1}^q\right),$$

Overview

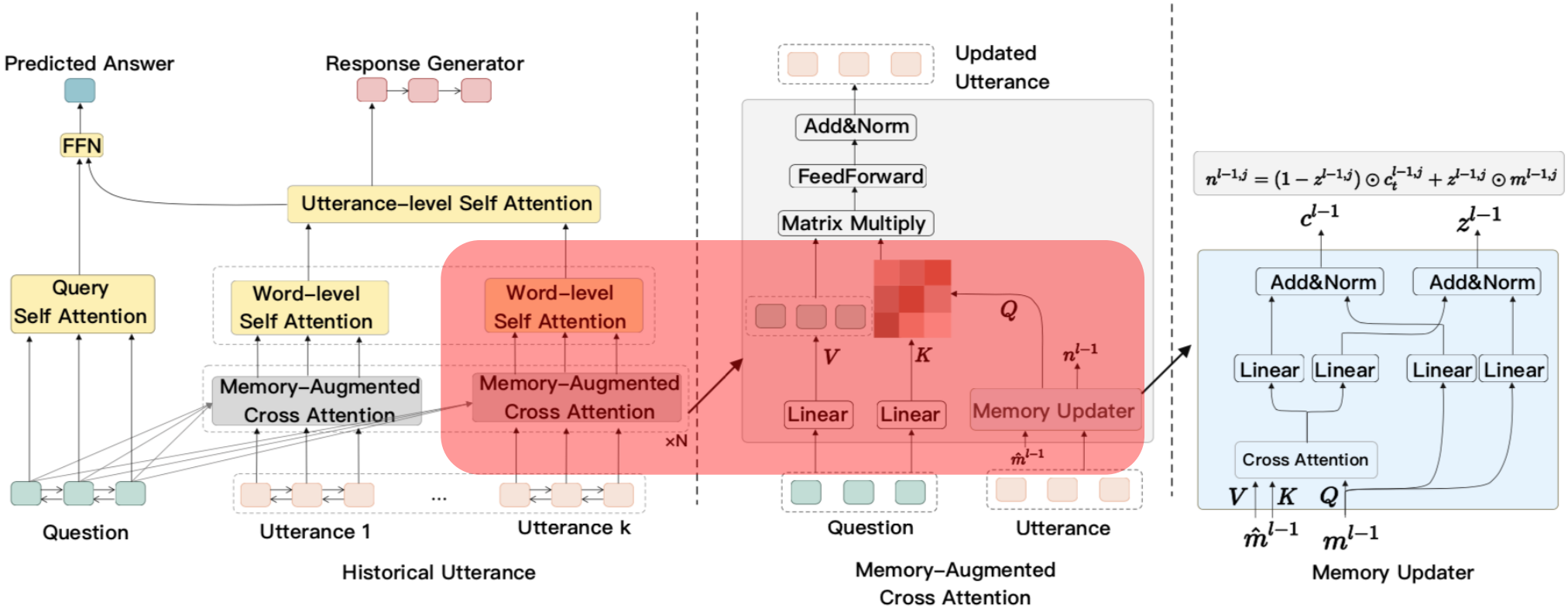
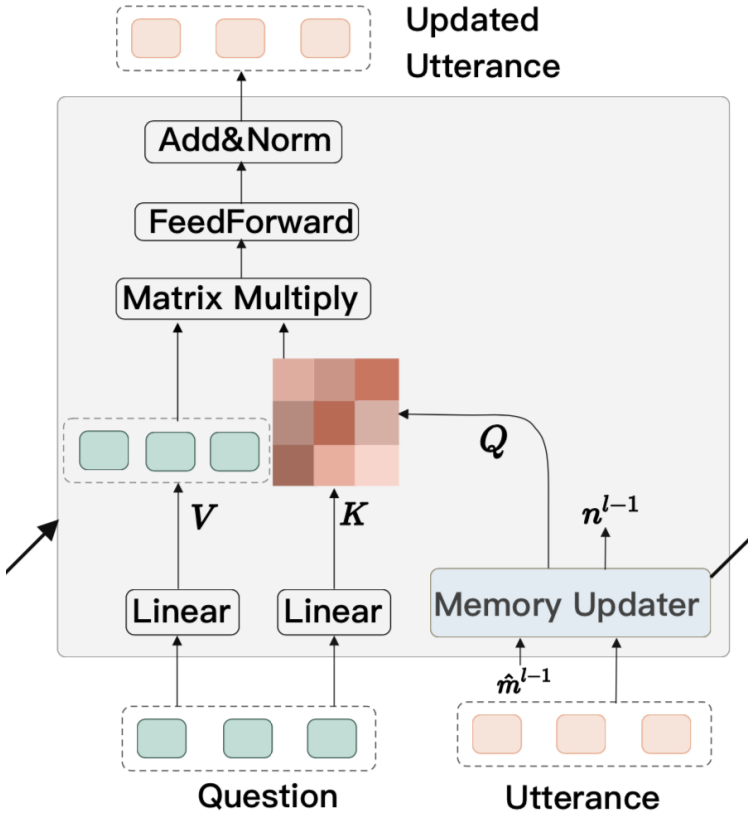


Figure 1: Overview of MRG. We divide our model into three parts: (1) Cross-hierarchical Encoder (which consists of memory-augment cross attention and two hierarchical self attentions); (2) Answer Selecter; (3) Response Generator.

Cross-Hierarchical Encoder

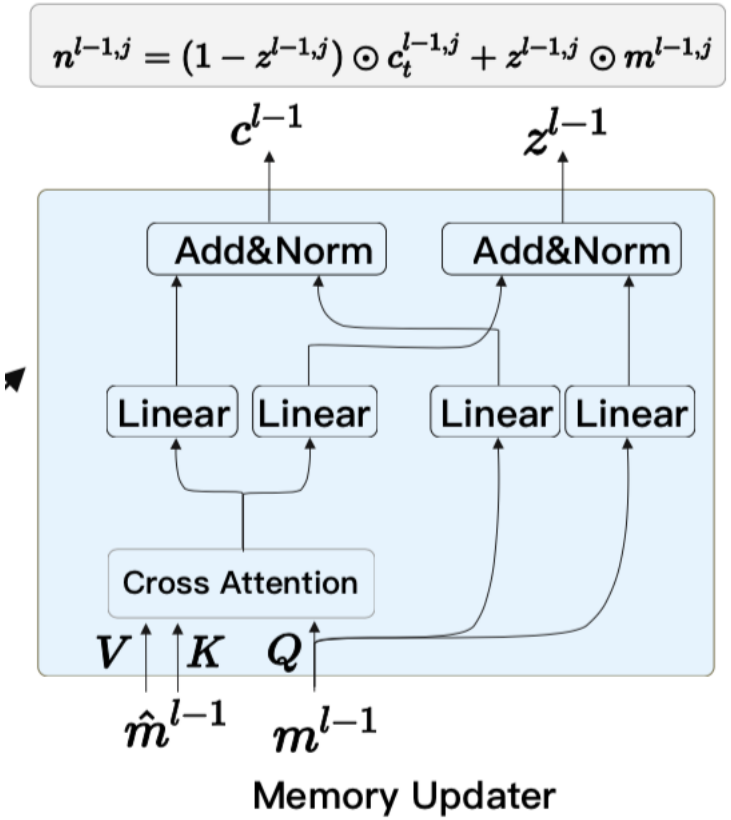
Memory augmented cross attention is based on the traditional Cross Attention Module (CAM) **Transformer architecture** (Vaswani et al. 2017).

While the aforementioned vanilla CAM is a powerful method, it is less suitable for multi-turn dialog due to its inability to fully utilize dialog history information. Thus, we augment it with an external memory module, which helps to remember and update history dialog information in a multi-slot way.



Cross-Hierarchical Encoder

$$
\begin{aligned}
s_i^{l-1,j} &= \mathrm{CAM}(m_i^{l-1,j}, \hat{m}_i^{l-1}), \\
c_i^{l-1,j} &= \tanh(W_a^{l-1} m_i^{l-1,j} + W_b^{l-1} s_i^{l-1,j}), \\
z_i^{l-1} &= \mathrm{sigmoid}(W_c^{l-1,j} m_i^{l-1,j} + W_d^{l-1} s_i^{l-1}), \\
n_i^{l-1,j} &= (1 - z_i^{l-1,j}) \odot c_t^{l-1,j} + z_i^{l-1,j} \odot m_i^{l-1,j},
\end{aligned}
$$



Overview

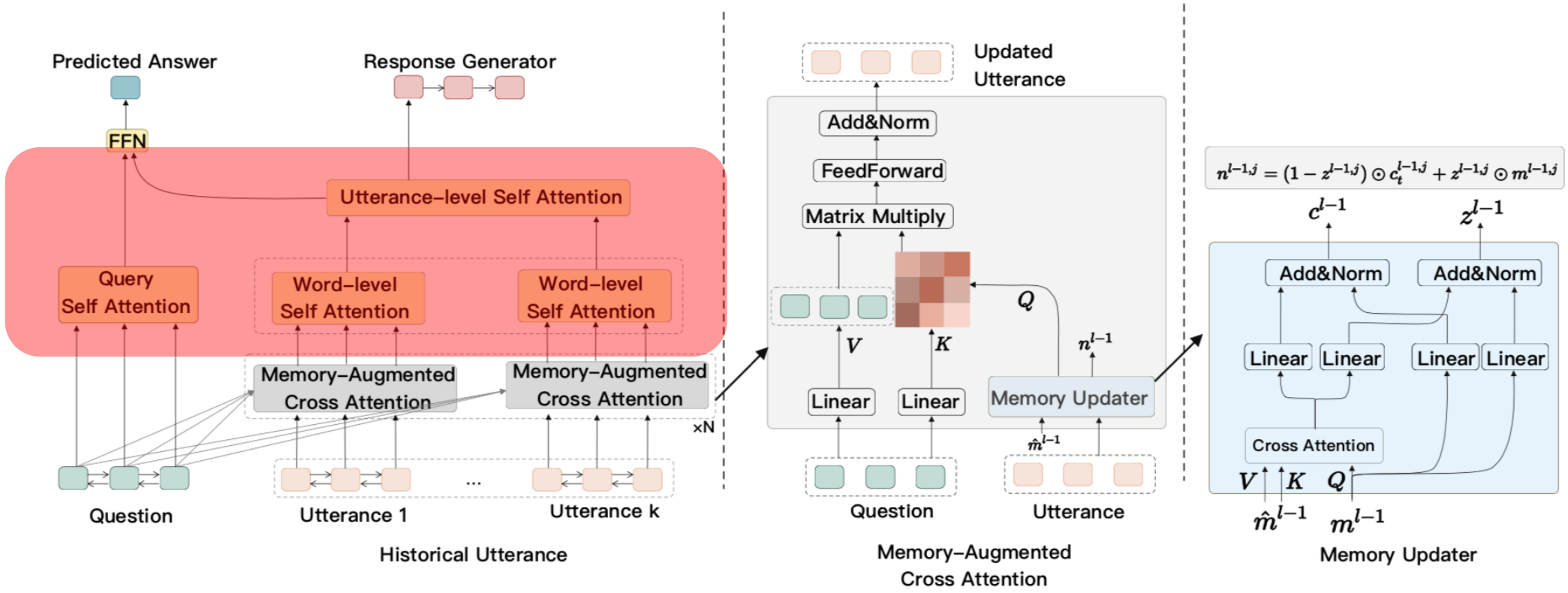
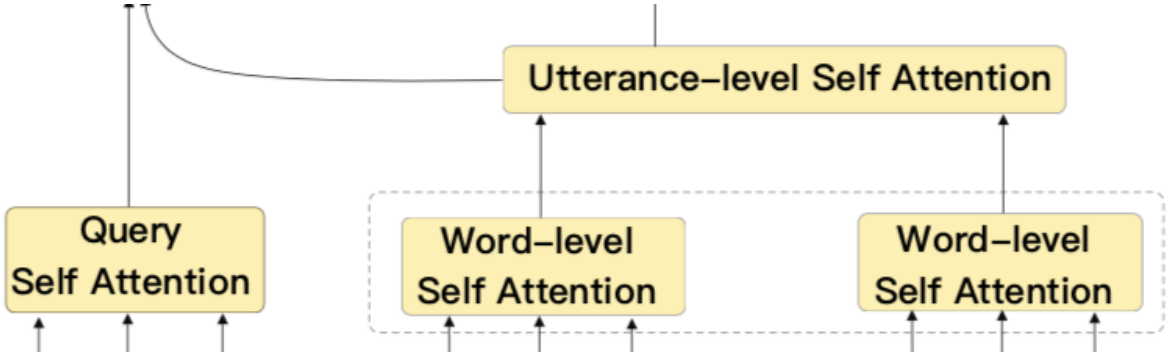


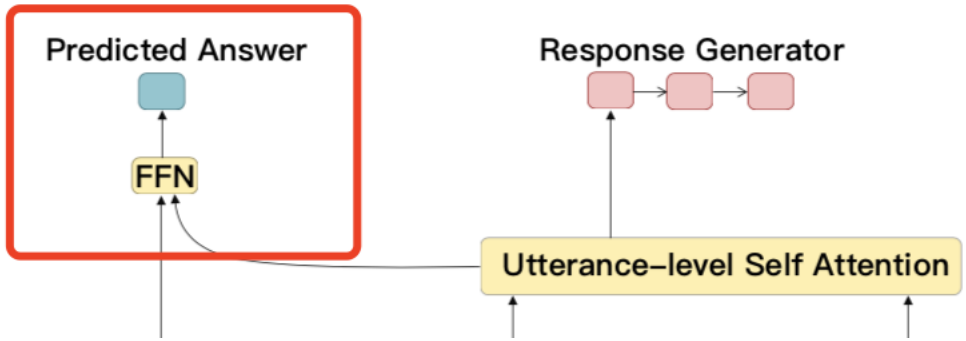
Figure 1: Overview of MRG. We divide our model into three parts: (1) Cross-hierarchical Encoder (which consists of memory-augment cross attention and two hierarchical self attentions); (2) Answer Selecter; (3) Response Generator.

Cross-Hierarchical Encoder

- Word-level
  - The first level in our hierarchical attention encodes each utterance independently from other utterances at word-level, resulting in a fixed-dimensional representation of each utterance
- Utterance-level
  - Similar to word-level attention, an utterance-level MAM is applied on these representations to fuse information between different utterances



Answer Selector



$$h^q = \text{meanpool}\left(\{h_1^q, \cdots, h_{N^q}^q\}\right),$$

$$\hat{A} = W_f \tanh\left(W_e\left[h^{u,1}; \ldots; h^{u,N^u}; h^q\right] + b^e\right) + b^f,$$

Response Generator

$$s_0 = W_g h^d + b_g,$$
$$s_t = \mathbf{LSTM}\left(s_{t-1}, [f_{t-1}; e(y_{t-1})]\right).$$

Dataset

- To our best knowledge, no existing works consider MRC in response generation task
- We propose a dialog reading comprehension dataset (DRCD)
- DRCD is based on the Restoration-200k dataset
  - The utterance with omitted information is manually annotated
- Such omitted information leads to a difficulty in fully understanding the dialog context and requires reasoning ability to for a model
- We hire an annotation team to write questions that are focused on the missing information
  - **A big THANK-YOU to Xiaomi Team**

Dataset

- Since it is time-consuming to write questions for the whole dataset, and based on the labeled answer it is rather easy to construct the question, we ask the team to write questions for 10k cases, and then automatically generate questions for the rest of the dataset
  - Concretely, we utilize PG to generate questions due to its good performance
- We conduct a human evaluation to examine the generation quality
  - The result shows that generated questions that score over 3 takes up 76.5%, showing that most of the generated questions are of good quality
- The kappa statistics indicate the moderate agreement between annotators

Dataset

- We randomly split the dataset with question-answer pair to 113,116 training, 3,000 validation, and 3,000 test cases
  - The average character-level context length and utterance length of the dataset is and 43.4 and 9.05
- Note that in the validation and test datasets the questions are all written by human, ensuring that the testing results are convincing

Experiment

- Results

Model	BLEU1	BLEU2	BLEU3	BLEU4	Average	Extrema	Greedy
Seq2Seq	0.2260	0.1566	0.0876	0.0671	0.4341	0.6695	0.7759
HRED	0.2273	0.1559	0.0871	0.0667	0.4320	0.6601	0.7885
VAE	0.2316	0.1586	0.0886	0.0680	0.4350	0.6396	0.7808
Transfomer	0.2181	0.1482	0.0825	0.0631	0.4407	0.6500	0.7920
PAC	0.2413	0.1624	0.0902	0.0689	0.4396	0.6447	0.7909
MRG	**0.2632**	**0.1735**	**0.0968**	0.0741	**0.4513**	**0.6769**	**0.8025**
MRG w/o MCAM	0.2224	0.1533	0.0857	0.0656	0.4436	0.6630	0.7837
MRG w/o MAM	0.2404	0.1616	0.0946	0.0665	0.4343	0.6740	0.7798
MRG w/o MemUpd	0.2498	0.1585	0.0894	**0.0747**	0.4419	0.6551	0.7884
MRG w/o MT	0.2231	0.1541	0.0862	0.0661	0.4343	0.6734	0.7645

Table 2: Automatic evaluation results on response generation task. The best results are bold.

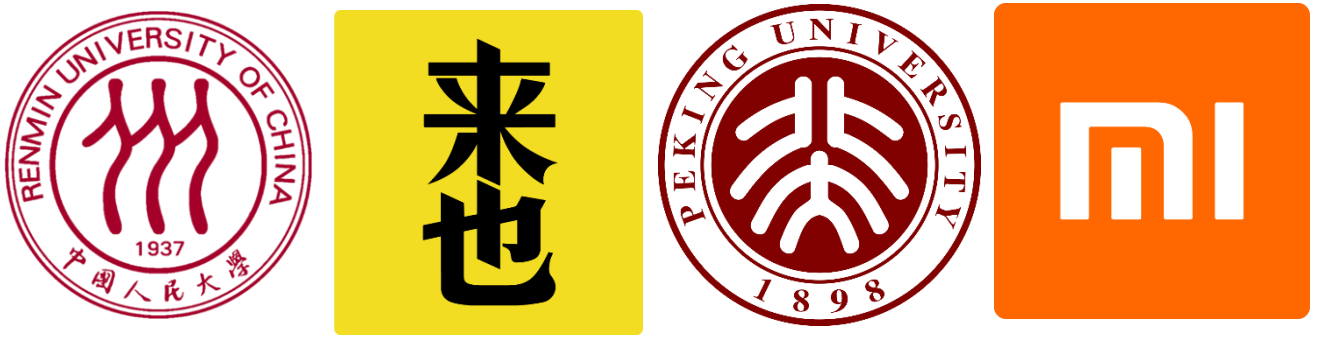
Case Study

	Example 1	Example 2	Example 3
A_1	有木有人带妹子吃喝玩乐在杭州 Is there anyone to take girls to eat, drink and have fun in Hangzhou	天蝎座不自恋真的就能死 Scorpio will die without narcissism	lz女坐标杭州心情不nice The author is a girl, located in Hangzhou, has a bad mood
B_1	没真相不敢带 I dare not bring a girl without a photo	挺准最近就是被一个天蝎虐 That's right. I've been abused by a Scorpio recently.	那怎么办 Then what to do
A_2	有了真相更不敢带哈 With a photo, you will dare not bring her more	嗯 Yes	睡觉么 What about sleep
B_2	犀利 Sharp	你是摩羯啊 So you are a Capricorn	也可以不过不是长久之计 That's good, but not a long-term solution
A_3	一般(犀利)啦 Generally (sharp)	(摩羯被天蝎虐)这就是宿命 This (Capricorn abused by Scorpio) is fate.	哈哈那(心情不nice)怎么办 Ha-ha, then what to do (if in bad mood)
Question	什么很一般 What is general	什么是宿命 What is fate	什么出问题了 What's wrong
Answer	犀利程度一般 The degree of sharpness	摩羯被天蝎虐 Capricorn abused by Scorpio	心情 Mood
Reference	一般都这么犀利男人咋办 If that is general sharp then what to do with man	所以我喜欢上天蝎就是个错误 So it was a mistake for me to like Scorpio	找个心灵的驿站 Find a post station for the soul
HRED	是你的错觉 It's your delusion	洗洗睡了 Wash and sleep	生命健康更加可贵 Life and health are more valuable
VAE	好吧,改天可以一起去嗨皮了 Well, we can play the other day	我帮同事发帖子勾搭上了 I helped my colleagues post and hook up	年龄大了就想睡 When you get older, you want to sleep
PAC	你也要去对不啦 So you will go there, aren't you?	那你还不如玩手机 You might as well play with your cell phone	我闲了就读书 I read when I'm free
MRG	是过于犀利啦 It's too sharp	还是要看人，怎么就宿命了哈哈哈 It still depends on people. You should not believe in fate lol.	出门旅游，逛逛豆瓣 Travel outside, visit Douban

Table 4: Responses generated by baselines and our model along with the QA pairs.

Summary

- A new dataset for reasoning in dialogs
- A framework for multi-tasking with new structures (memory)
  - Machine reading comprehension
  - Dialog generation
- Effectiveness in experiments verified



Multilingual COVID-QA:

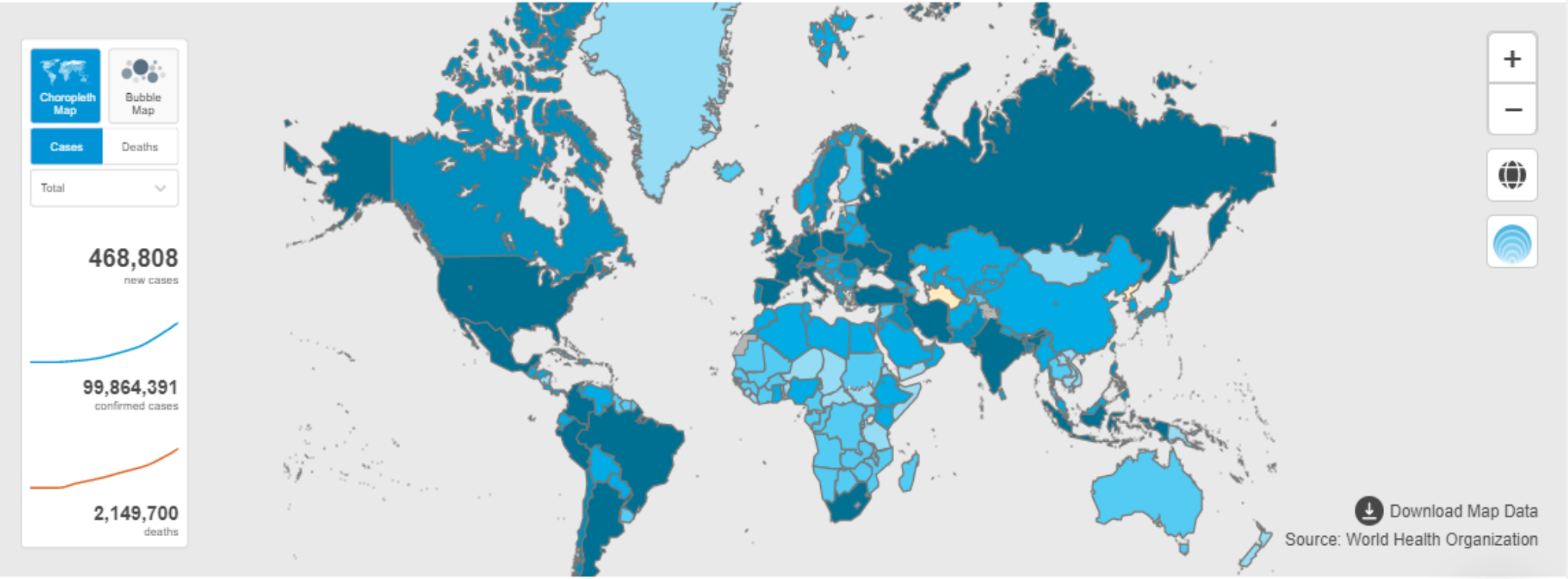
Learning towards Global Information Sharing via Web Question Answering in Multiple Languages

Rui Yan, Weiheng Liao, Jianwei Cui, Hailei Zhang, Yichuan Hu and Dongyan Zhao

WWW 2021

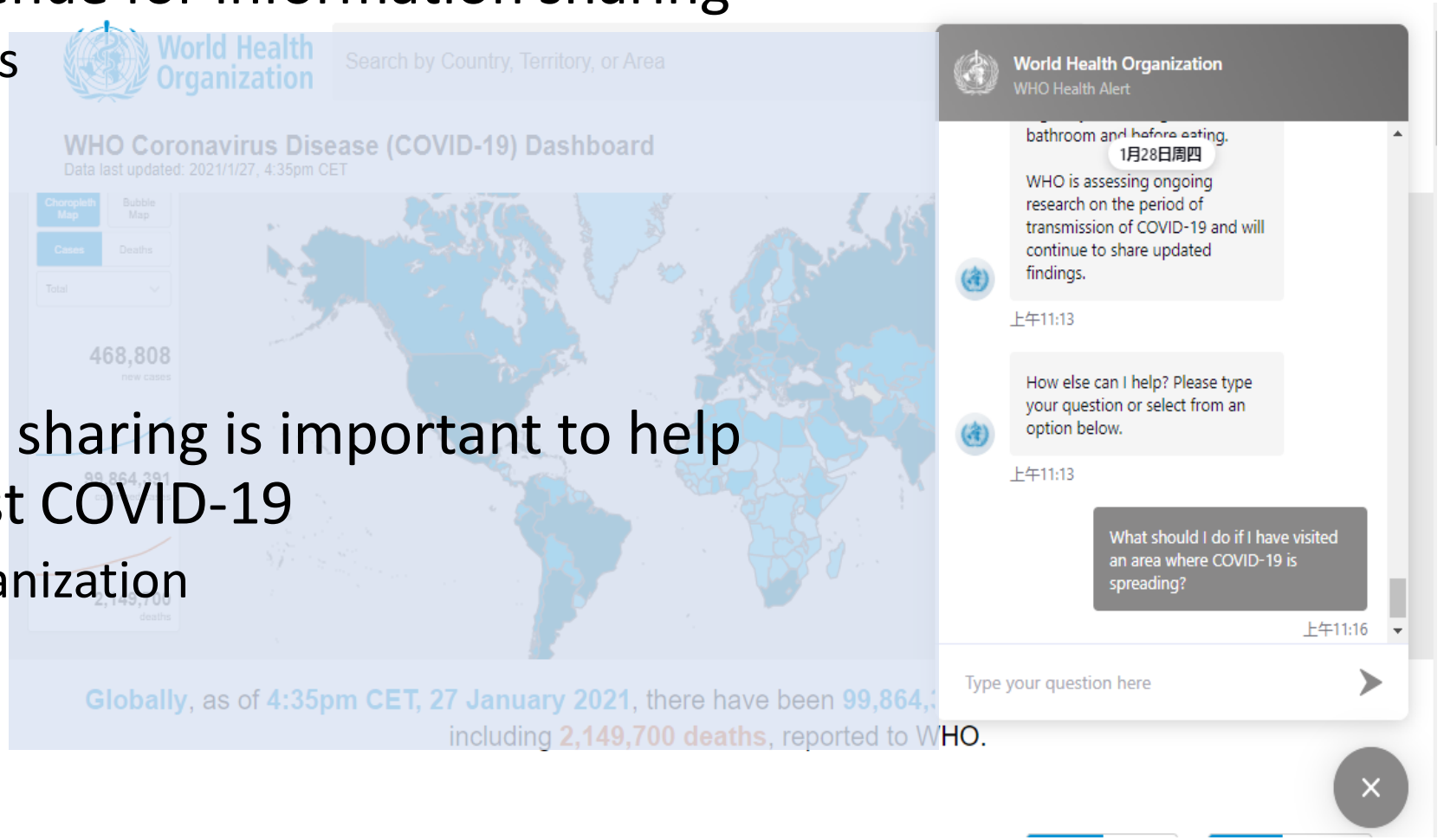
COVID-19!!

- Global pandemic: global policy



Why Web QA

- Web is an active venue for information sharing
  - News and bulletins
  - FAQ
  - Social media
  - Search engines
  - QA forums
- Global information sharing is important to help people fight against COVID-19
  - World Health Organization



Challenges

- Given that information is conveyed in different languages
  - Unfriendly to monolingual speakers
- Insufficient information for a particular language
  - Especially for a low-resource language
- Translation is not ALWAYS reliable
  - Especially for a specific domain such as COVID-19
- Information in different languages is not ALWAYS aligned
  - E.g. to share experiences in Chinese which is not described in other languages

Contribution

- Information collection from the global data and utilize knowledge to reinforce between one language and another
- Improving translation models extended from the general domain
- Unsupervised language alignment and cross-lingual mapping for the non-parallel data

Preliminary

- Data collection
  - (q,a) pairs in different languages, i.e., English, Chinese and Japanese
    - Not always in parallel
  - Large-scale language model pre-training
- Unsupervised NMT
  - Word-to-word language alignment
  - Weak translation models from one language to another
- Encoder-decoder
  - Bi-GRU
  - Other structures may apply, too

Formulation

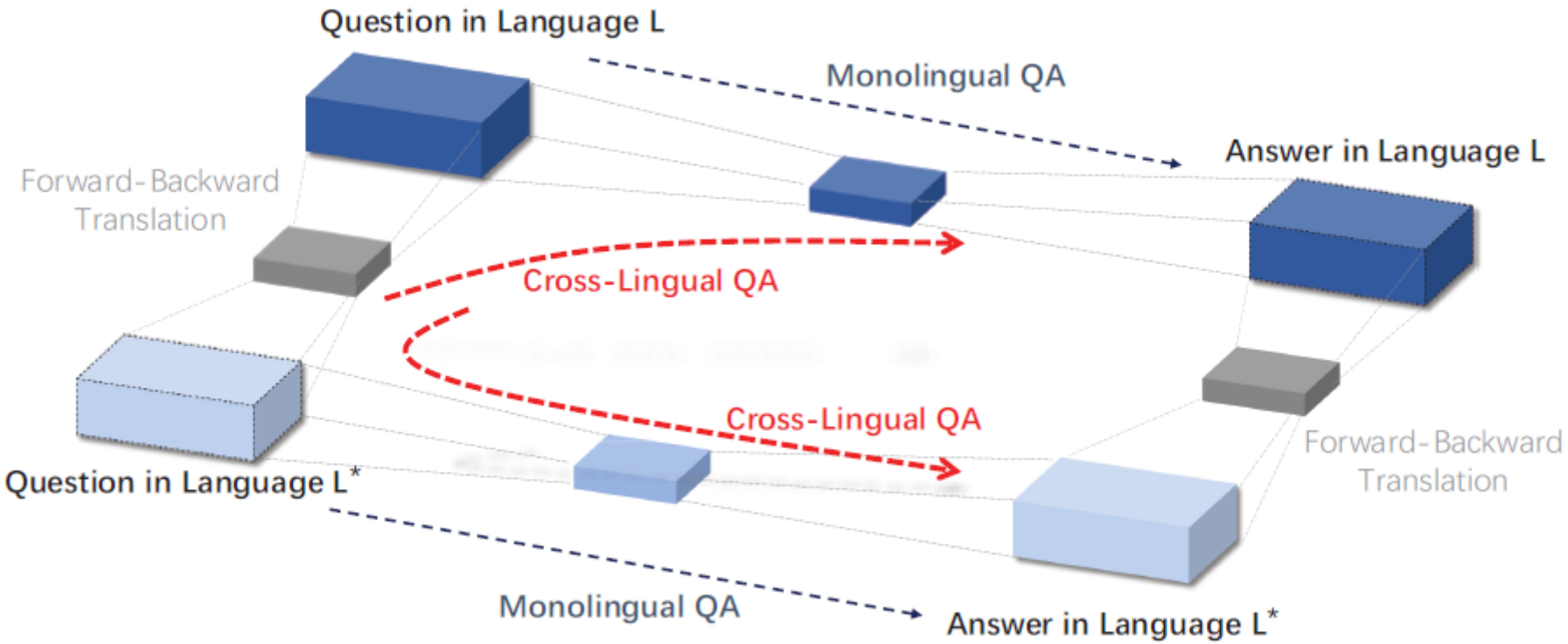
- Dataset $D^l = \{(q,a)\}^l$ and $D^{l^*} = \{(q,a)\}^{l^*}$
- There can be a small subset of parallel data $P=\{(q^l, a^{l^*})\}$
- To generate an answer in the same language
  - Information is mixed across languages

$$p(a^\ell | q^\ell) = \prod_{t=1} p(a_t^\ell | q^\ell, a_{<t}^\ell).$$

  - Two functions will be learned g(.) and f(.), for generation and translation respectively

Multilingual QA

- Framework Overview



QA Translation

- Forward translation and backward translation
  - From one language to another, and then translation back in a dual process
  - Expectation: perfectly reconstruct the original sentence

$$\mathcal{L}_1 = -\log p(q^\ell | f^{\ell \leftarrow \ell^\star}(f^{\ell \rightarrow \ell^\star}(q^\ell)))$$

$$\mathcal{L}_2 = -\log p(a^\ell | f^{\ell \leftarrow \ell^\star}(f^{\ell \rightarrow \ell^\star}(a^\ell)))$$

- Translation loss

$$\mathcal{L}_{\text{trans}} = \mathcal{L}_1 + \mathcal{L}_2$$

QA Generation

- Monolingual QA
  - Unconditioned on translation models

$$\mathcal{L}_3 = -\log p(a^\ell | g_{\text{mono}}(q^\ell))$$

  - Conditioned on translation models

$$\mathcal{L}_4 = -\log p(f^{\ell \leftarrow \ell^\star}(f^{\ell \rightarrow \ell^\star}(a^\ell)) | g_{\text{mono}}(f^{\ell \leftarrow \ell^\star}(f^{\ell \rightarrow \ell^\star}(q^\ell))))$$

- Monolingual QA loss

$$\mathcal{L}_{\text{mono}} = \mathcal{L}_3 + \mathcal{L}_4$$

Cross-Lingual QA

- Knowledge can be shared through the semantics
  - Translate the question part

$$\mathcal{L}_5 = -\log p(a^{\ell} | g_{\text{cross}}(f^{\ell \to \ell^{\star}}(q^{\ell})))$$

  - Translate the answer part

$$\mathcal{L}_6 = -\log p(f^{\ell \to \ell^{\star}}(a^{\ell}) | g_{\text{cross}}(q^{\ell}))$$

- When parallel data available

$$\mathcal{L}'_5 = \log(a^{\ell^{\star}} | g_{\text{cross}}(q^{\ell}))$$

$$\mathcal{L}'_6 = \log(a^{\ell} | g_{\text{cross}}(q^{\ell^{\star}}))$$

- Cross-lingual QA loss

$$\mathcal{L}_{\text{cross}} = \begin{cases} \mathcal{L}_5 + \mathcal{L}_6 & \text{(non-parallel data)} \\ \mathcal{L}'_5 + \mathcal{L}'_6 & \text{(parallel data)} \end{cases}$$

Translation-Generation Joint Learning

- Training steps
  - Language model pretraining
  - Monolingual QA generation pretraining
  - Fine-tune parameters on the parallel dataset and pseudo parallel dataset
- Final objective

$$\mathcal{L}_{\ell} = \mathcal{L}_{\text{trans}} + L_{\text{mono}} + \mathcal{L}_{\text{cross}}$$

- Training two languages in pairs

$$\mathcal{L} = \mathcal{L}_{\ell} + \mathcal{L}_{\ell^{\star}}$$

Datasets

- Information hubs: FAQ
- Expert interview QA
- Web QA forums

	English	Chinese	Japanese
# of Questions	52,006	39,231	11,393
# of Answers	91,752	73,043	25,012
# of Pretraining Sentences	500,000	500,000	500,000
# of Vocabulary	83,520	101,529	45,188
Avg. Question Length	28.85	31.26	33.69
Avg. Answer Length	158.82	128.26	187.18
# Training QA pairs	68,425	48,369	14,500
# Validation QA Pairs	8,555	6,050	1,810
# Testing QA Pairs	8,555	6,050	1,810

Evaluation Metric

- BLEU
- Semantic similarity via embeddings
  - Greedy, Extrema, Average
- Human evaluation
  - Non-experts: ordinary users
  - 500 test cases

Comparison Methods

- Monolingual QA
  - Train models separately
- Multi-tasking QA
  - Multilingual encoder-decoder jointly
- Translation-Aided
  - Using external translation models: Google Translate
  - Data-level
- Memory-Shared Model
  - Shared-private memory for different languages

Overall Performance

- Results for English and Chinese

	BLEU				EMBEDDING SIMILARITY			RATING
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Greedy	Average	Extrema	Human
Monolingual	6.115	0.518	0.292	0.188	0.725	0.628	0.301	1.008
Multi-Task	7.711	0.885	0.307	0.165	0.738	0.632	0.318	1.235
Transfer	6.331	0.716	0.258	0.159	0.721	0.614	0.272	1.131
Translation-Aided	4.725	0.128	0.015	0.006	0.623	0.619	0.211	1.158
Memory	8.700	1.035	0.337	0.189	0.777	0.652	0.358	1.392
Multilingual (Ours)	**9.925**	**1.502**	**0.610**	**0.331**	**0.805**	**0.694**	**0.376**	**1.519**

	BLEU				EMBEDDING SIMILARITY			RATING
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Greedy	Average	Extrema	Human
Monolingual	7.232	0.717	0.133	0.035	0.521	0.711	0.391	0.898
Multi-Task	7.802	0.783	0.125	0.025	0.529	0.703	0.413	0.925
Transfer	8.295	0.630	0.121	0.027	0.569	0.678	0.361	0.909
Translation-Aided	8.507	0.679	0.084	0.018	0.525	0.716	0.394	0.901
Memory	7.516	0.838	0.182	0.061	0.580	0.711	0.400	1.113
Multilingual (Ours)	**8.896**	**0.913**	**0.201**	**0.063**	**0.592**	**0.727**	**0.419**	**1.362**

Overall Performance

- Results for Japanese

	BLEU				EMBEDDING SIMILARITY			RATING
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Greedy	Average	Extrema	Human
Monolingual	5.100	0.533	0.229	0.158	0.583	0.487	0.326	0.741
Multi-Task	5.171	0.536	0.231	**0.189**	0.611	0.501	0.344	0.755
Transfer	5.041	0.530	0.247	0.187	0.595	0.508	0.345	0.749
Translation-Aided	6.148	0.582	0.138	0.033	0.640	0.527	0.365	0.761
Memory	6.315	0.718	0.233	0.106	0.622	0.519	0.366	0.958
Multilingual (Ours)	**6.320**	**0.725**	**0.253**	0.168	**0.659**	**0.539**	**0.371**	**1.096**

Ablation Results

- Components
  - Monolingual pretraining
  - Word-to-word weak translation
  - Forward-backward translation
  - Monolingual mapping
  - Cross-lingual mapping

	BLEU				EMBEDDING SIMILARITY		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Greedy	Average	Extrema
w/o pretraining	3.305	0.116	0.011	0.003	0.412	0.481	0.177
w/o word translation	3.947	0.132	0.021	0.005	0.435	0.492	0.185
w/o f-b translation	5.326	0.523	0.261	0.137	0.607	0.520	0.321
w/o mono mapping	1.653	0.082	0.007	0.001	0.343	0.295	0.106
w/o cross mapping	6.798	0.952	0.270	0.160	0.622	0.579	0.258
Full Model	8.380	1.047	0.355	0.187	0.685	0.653	0.389

Additional Analysis

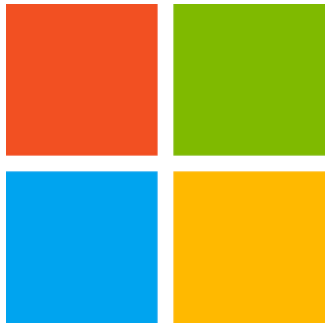
- Parallel Data
  - A small subset of parallel data vs. pseudo parallel data only

	BLEU				EMBEDDING SIMILARITY		
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Greedy	Average	Extrema
w/o Parallel Data	8.375	1.044	0.353	0.181	0.682	0.649	0.388
w/ Parallel Data	8.380	1.047	0.355	0.187	0.685	0.653	0.389

  - It is helpful to use the true parallel data for calibration
- Connection with existing models
  - Monolingual QA generation without the cross-lingual part
  - Backward-forward translation without QA generation: Dual Learning

Summary

- An effort from computer scientists to combat COVID-19
- A new framework for multilingual QA
  - Improved translation models
  - Improved monolingual QA models
  - Improved cross-lingual alignment and mapping
- Good performance in practice



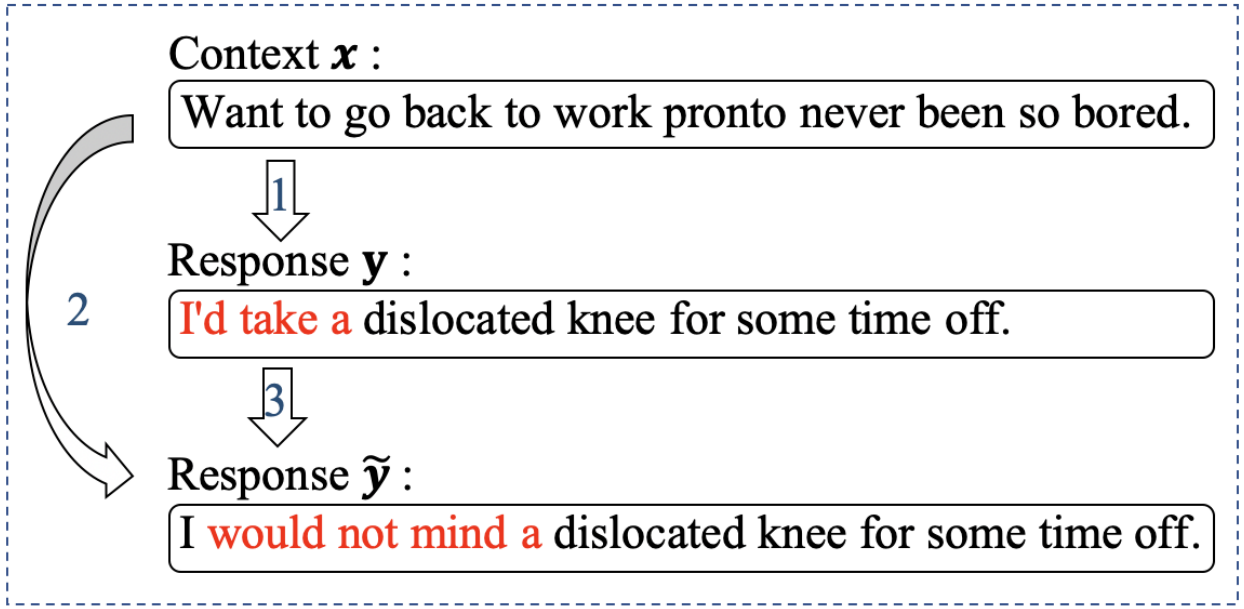
Stylized Dialogue Generation with Multi-Pass Dual Learning

Jinpeng Li, Yingce Xia, Hongda Sun, Dongyan Zhao, Tie-Yan Liu and **Rui Yan**

NeurIPS 2021

Movitation

- **Stylized dialogue generation**, which aims to generate a given-style response for an input context, plays a vital role in intelligent dialogue systems.



Example

Movitation

- There is no parallel data between the contexts and the responses of target style S_1, existing works mainly use back translation to generate stylized synthetic data for training, where the data about context, target style S_1 and an intermediate style S_0 is used.
- However, the interaction among these texts is not fully exploited, and the pseudo contexts are not adequately modeled.

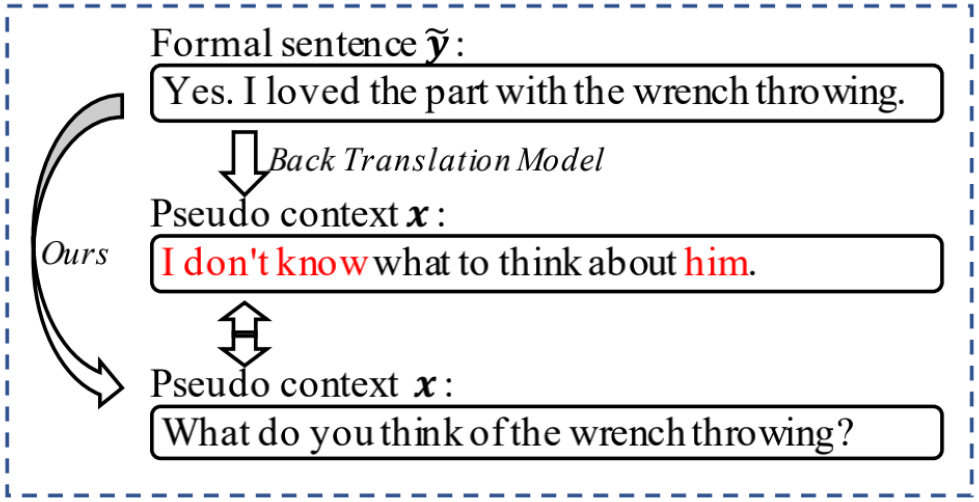


Figure 1: Different models produce pseudo context.

Movitation

- We propose multi-pass dual learning (MPDL), which leverages the duality among the context, response of style S_1 and response of style S_0.
- MPDL builds mappings among the above three domains, where the context should be reconstructed by the MPDL framework.
- We also introduce two discriminators to evaluate the quality of the generated data.

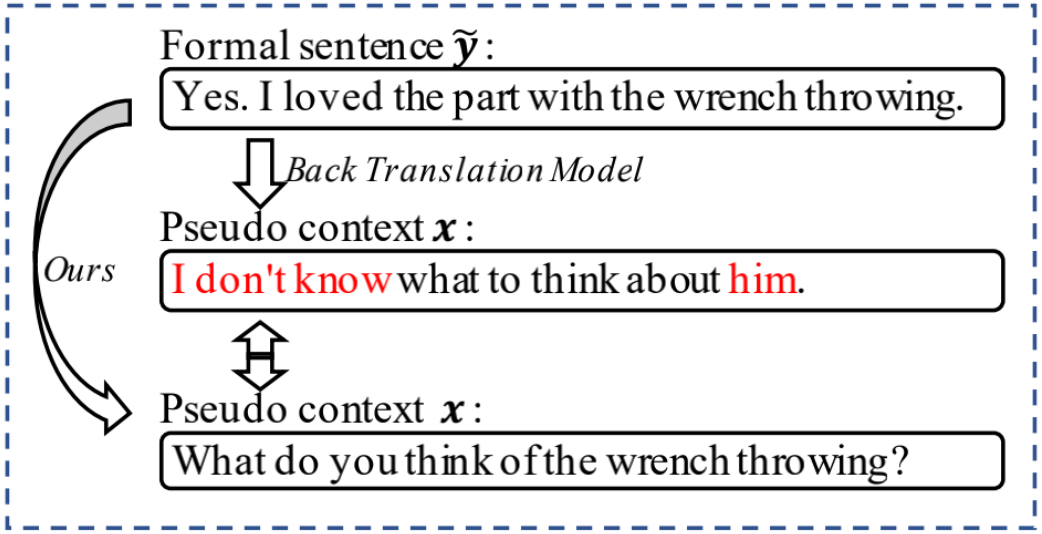


Figure 1: Different models produce pseudo context.

Contribution

- We propose multi-pass dual learning (MPDL) framework for stylized dialogue response generation, that can effectively leverage the unlabeled data;
- Compared with standard dual learning, we introduce two discriminators to evaluate the quality of the pseudo parallel data. This is a new attempt for the general dual learning framework;
- We provide a new dataset for this task and set several benchmarks using our method;
- We empirically verify the effectiveness of MPDL on two datasets with formal and Shakespearean response generation.

Multi-pass dual learning framework

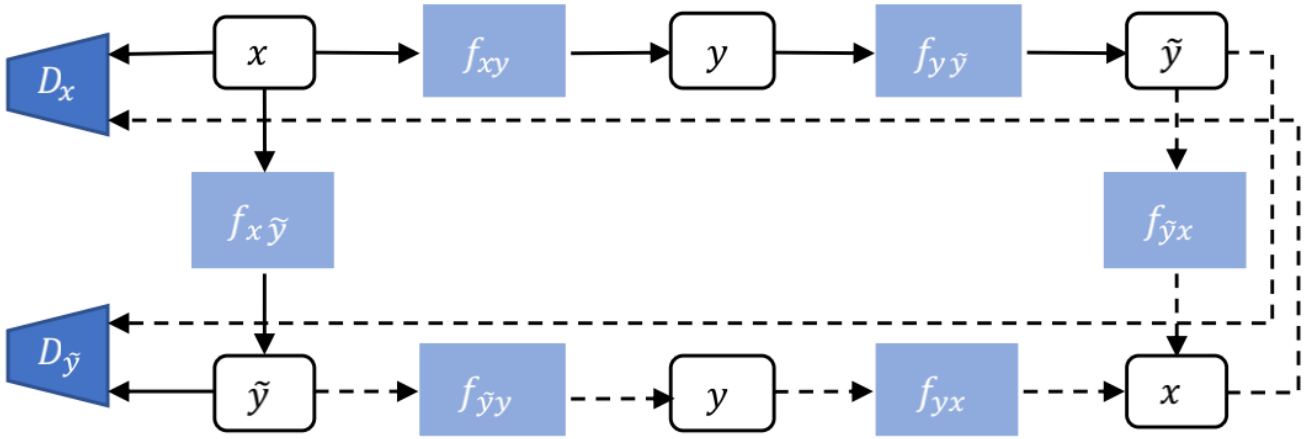


Figure 1: The framework multi-pass dual learning. f denotes the encoder and decoder model constructed by GPT-2. x, y and $\tilde{y}$ denote context, style S_0 and style S_1 respectively. D_x and $D_{\tilde{y}}$ are discriminators.

There are three pairs of dual tasks involved in our framework.

$$f_{xy} : \mathcal{C} \mapsto \mathbb{S}_0, f_{yx} : \mathbb{S}_0 \mapsto \mathcal{C}$$

Model Architecture

- Three pairs of dual tasks：
  - Dialogue generation and the inversed task $f_{xy} : \mathcal{C} \mapsto \mathbb{S}_0, f_{yx} : \mathbb{S}_0 \mapsto \mathcal{C}$
  - Style transfer between S0 and S1 $f_{y\tilde{y}} : \mathbb{S}_0 \mapsto \mathbb{S}_1, f_{\tilde{y}y} : \mathbb{S}_1 \mapsto \mathbb{S}_0$
  - Stylized dialogue response generation and the inversed task $f_{x\tilde{y}} : \mathcal{C} \mapsto \mathbb{S}_1, f_{\tilde{y}x} : \mathbb{S}_1 \mapsto \mathcal{C}$

- Goal: $f_{x\tilde{y}}$, Max: $\log P(\tilde{y}' = \tilde{y}|\tilde{y}; f_{\tilde{y}y}, f_{yx}, f_{x\tilde{y}})$.

$$\log P(\tilde{y}' = \tilde{y}|\tilde{y}; f_{\tilde{y}y}, f_{yx}, f_{x\tilde{y}}) = \log \sum_{x \in \mathcal{C}} \sum_{y \in \mathbb{S}_0} P(\tilde{y}', y, x|\tilde{y}; f_{\tilde{y}y}, f_{yx}, f_{x\tilde{y}})$$

$$= \log \sum_{x \in \mathcal{C}} \sum_{y \in \mathbb{S}_0} P(\tilde{y}'|y, x, \tilde{y}; f_{\tilde{y}y}, f_{yx}, f_{x\tilde{y}}) P(y, x|\tilde{y}; f_{\tilde{y}y}, f_{yx}, f_{x\tilde{y}}).$$

$$\log P(\tilde{y}' = \tilde{y}|\tilde{y}; f_{\tilde{y}y}, f_{yx}, f_{x\tilde{y}}) \geq \sum_{x \in \mathcal{C}} \sum_{y \in \mathbb{S}_0} P(y, x|\tilde{y}; f_{\tilde{y}y}, f_{yx}, f_{x\tilde{y}}) \log P(\tilde{y}'|x, f_{x\tilde{y}})$$

$$= \sum_{x \in \mathcal{C}} \sum_{y \in \mathbb{S}_0} P(x|y; f_{yx}) P(y|\tilde{y}; f_{\tilde{y}y}) \log P(\tilde{y}'|x, f_{x\tilde{y}}),$$

Model Architecture

- Based on Transformer:
  - Initialized using pretrained DialoGPT weights
  - The parameters of all forward models are shared
  - The parameters of all backward models are shared
  - The parameters of encoder and decoder are shared

$$H_f(x) = \texttt{enc}_f(E_T + \mathcal{W}(x))$$
$$H_f^{\xi}(x) = E_{\xi} + H_f(x),$$

Model Architecture

- Training:
  - Given an input x and the task T, the encoder of the forward model works as follows:

$$H_f(x) = \mathrm{enc}_f(E_T + \mathcal{W}(x))$$

$$H_f^{\xi}(x) = E_{\xi} + H_f(x),$$

  - $H_f^{s_0}(x)$ for dialogue response generation, $H_f^{s_1}(x)$ for stylized dialogue generation
  - $H_f^{s_1}(y)$ for text style transfer, $H_b^{s_0}(y)$ for inverse dialogue response generation
  - $H_b^{s_1}(\tilde{y})$ for inverse stylized dialogue generation and inverse text style transfer.

Model Architecture

- Loss:

$$\mathcal{L}_{\text{dia}} = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\text{dia}}} -(\log P(y|x; f_{xy}) + \log P(x|y; f_{yx})),$$

$$\mathcal{L}_{\text{tra}} = \mathop{\mathbb{E}}_{(y,\tilde{y})\sim\mathcal{D}_{\text{tra}}} -(\log P(\tilde{y}|y; f_{y\tilde{y}}) + \log P(y|\tilde{y}; f_{\tilde{y}y})),$$

$$\mathcal{L}_{\text{dual},1}(\tilde{y}) = -(\log P(\tilde{y}|x'; f_{x\tilde{y}})),$$

$$\mathcal{L}_{\text{dual},2}(x) = -(\log P(x|\tilde{y}'; f_{\tilde{y}x})),$$

Quality Discriminator

- Previous work cannot obtain proper context sequences.

Examples include:

```
I don't know ...
I'm so happy ...
I'm not sure if I should ...
```

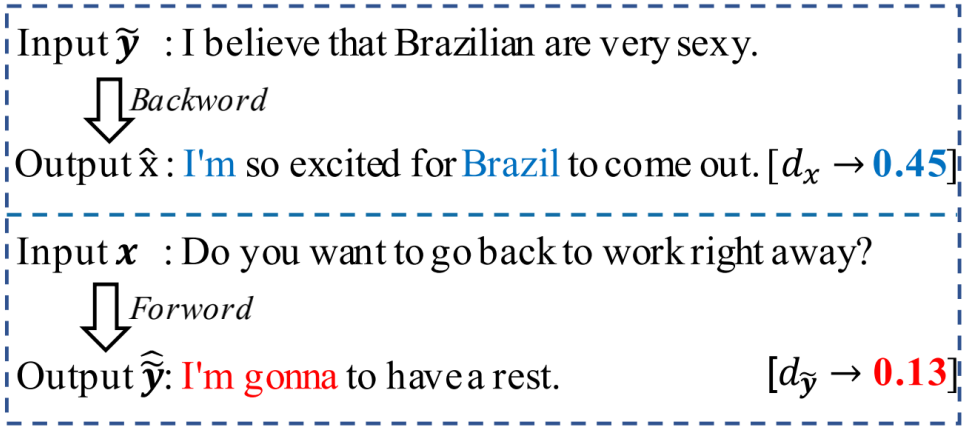


Figure 3: The discriminators of our proposed.

$$\ell^x_{\text{dis}} = \log D_x(H_f^{S_0}(x)) + \log(1 - D_x(H_f^{S_1}(y))),$$
$$\ell^{\tilde{y}}_{\text{dis}} = \log D_{\tilde{y}}(H_b^{S_1}(\tilde{y})) + \log(1 - D_{\tilde{y}}(H_b^{S_0}(y))),$$
$$\mathcal{L}_{\text{dis}} = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\text{dia}};(\tilde{y})\sim\mathcal{D}_{\text{sty}}} -(\ell^x_{\text{dis}} + \ell^{\tilde{y}}_{\text{dis}}).$$

$$\mathcal{L}^*_{\text{dual},1} = \mathop{\mathbb{E}}_{\tilde{y}\sim\mathcal{D}_{\text{sty}}} - D_x(H_f^{S_0}(x')) \log P(\tilde{y}|x'; f_{x\tilde{y}}),$$

$$\mathcal{L}^*_{\text{dual},2} = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\text{dia}}} - D_{\tilde{y}}(H_b^{S_1}(\tilde{y}')) \log P(x|\tilde{y}'; f_{\tilde{y}x}),$$

Algorithm

- Mixed Loss:

$$\mathcal{L} = \mathcal{L}_{\text{dia}} + \mathcal{L}_{\text{tra}} + \mathcal{L}^*_{\text{dual},1} + \mathcal{L}^*_{\text{dual},2}.$$

Algorithm 1 The training process

Require: Parallel data $\mathcal{D}_{\text{dia}}$, $\mathcal{D}_{\text{tra}}$, unpair data $\mathcal{D}_{\text{sty}}$ and parameters $\overrightarrow{\Theta}$, $\overleftarrow{\Theta}$
Ensure: The MPDL model $f_{x\tilde{y}}$

Initialize the forward and backward encoder-decoder parameters $\overrightarrow{\Theta}$, $\overleftarrow{\Theta}$ using DialoGPT
Define $N \leftarrow$ freeze steps
repeat
  Sample b mini-batch size dialogue pairs $\mathcal{D}^b_{\text{dia}} \subset \mathcal{D}_{\text{dia}}$
  Sample b mini-batch size transfer pairs $\mathcal{D}^b_{\text{tra}} \subset \mathcal{D}_{\text{tra}}$
  Train f_{xy}, f_{yx} by obtaining Loss $\mathcal{L}_{\text{dia}}$ on $\mathcal{D}^b_{\text{dia}}$
  Train $f_{y\tilde{y}}$, $f_{\tilde{y}y}$ by obtaining Loss $\mathcal{L}_{\text{tra}}$ on $\mathcal{D}^b_{\text{tra}}$
  if Current Step $> N$ **then**
    Sample b mini-batch size style sentences $\mathcal{D}^b_{\text{sty}} \subset \mathcal{D}_{\text{sty}}$ for backward stage
    Sample b mini-batch size contexts $\mathcal{D}^b_{\text{dia}} \subset \mathcal{D}_{\text{dia}}$ for forward stage
    Decode the pseudo context x' and $\tilde{y}'$ from $f_{yx}(f_{\tilde{y}y}(\tilde{y}))$ and $f_{y\tilde{y}}(f_{xy}(x))$
    Train $f_{x\tilde{y}}$, $f_{\tilde{y}x}$ by obtaining Loss $\mathcal{L}^*_{\text{dual},1}$, $\mathcal{L}^*_{\text{dual},2}$ on $\{x', \tilde{y}\}$, $\{x, \tilde{y}'\}$
  end if
  Optimizing MPDL model with the mixed Loss $\mathcal{L}$
until The model converge
return $f_{x\tilde{y}}$

Datasets

- TCFC: informal to formal
- Shakespearean Dialogue Generation Corpus (SDGC): modern to Shakespearean

Table 5: The statistics of datasets.

Datasets		Train	Valid	Test
TCFC	$\mathcal{D}_{\text{dia}}$	217,222	978	978
	$\mathcal{D}_{\text{sty}}$	500,000		
	Avg.	12.44	14.49	14.49
GYAFC	E&M	52,595	2,877	1,416
	F&R	51,967	2,788	1,332
	Avg.	10.69	10.39	10.71
SDGC	$\mathcal{D}_{\text{dia}}$	217,222	1000	1000
	$\mathcal{D}_{\text{sty}}$	18,395		
	Avg.	12.07	9.12	9.12
Shakespeare	$\mathcal{D}_{\text{sty}}$	18,395	1,218	1,462
	Avg.	11.38	10.47	7.77

Experimentsal Results

Table 1: Automatic and manual evaluation results on TCFC dataset.

Model	Automatic Metrics					Manual Metrics		
	BLEU-1	BLEU-2	Distinct	BERT	SVM	Fluency	Relevance	Style-Con.
The target style is formal response (i.e., style S_1)								
MTask	6.35	0.50	29.3	37.3	50.1	0.78	0.33	0.58
SLM	12.6	0.99	42.5	85.6	87.2	0.83	0.45	0.87
SFusion	5.51	0.28	**61.0**	21.9	39.0	0.77	0.32	0.57
S2S+BT	12.1	1.25	42.0	86.3	86.8	0.79	0.31	0.65
SRJT	15.1	1.71	43.4	97.3	96.1	0.85	0.55	0.89
MPDL	**16.5**	**2.07**	51.3	**98.6**	**97.1**	**0.88**	**0.64**	**0.91**
Human	-	-	62.7	89.6	85.8	0.88	0.65	0.90
The target style is informal response (i.e., style S_0)								
S2S	6.92	0.61	54.8	70.1	60.9	0.75	0.46	0.66
SFusion	4.61	0.22	**62.8**	**70.3**	**61.1**	0.66	0.34	**0.78**
SRJT	6.96	0.67	49.4	69.4	59.2	0.81	**0.57**	0.73
MPDL	**7.12**	**0.69**	49.5	**70.3**	60.7	**0.83**	0.55	0.75
Human	-	-	72.6	72.0	72.1	0.79	0.56	0.78

Experimentsal Results

Table 2: Automatic and manual evaluation results on SDGC dataset.

Model	Automatic Metrics					Manual Metrics		
	BLEU-1	BLEU-2	Distinct	BERT	SVM	Fluency	Relevance	Style-Con.
The target style is Shakespearean (i.e., style S_1)								
MTask	5.64	0.15	30.2	18.4	20.7	0.45	0.23	0.19
SFusion	6.76	0.21	41.6	20.1	16.0	0.50	0.27	0.21
S2S+BT	9.68	0.38	37.4	**84.7**	**73.0**	0.64	0.33	0.70
SRJT	12.2	0.89	43.1	65.0	53.4	0.67	0.59	0.71
MPDL	**13.9**	**1.43**	**53.2**	64.3	54.1	**0.79**	**0.71**	**0.80**
Human	-	-	81.2	92.9	66.0	0.84	0.75	0.83
The target style is modern English (i.e., style S_0)								
S2S	9.33	0.90	43.2	93.1	94.3	0.64	0.31	0.55
SFusion	5.71	0.12	45.7	91.5	93.9	0.45	0.41	0.63
SRJT	8.90	0.63	40.5	96.9	95.4	0.71	0.54	0.64
MPDL	**10.4**	**1.19**	**46.7**	**97.8**	**96.7**	**0.73**	**0.57**	**0.70**
Human	-	-	79.9	92.9	92.0	0.80	0.69	0.77

Experimentsal Results

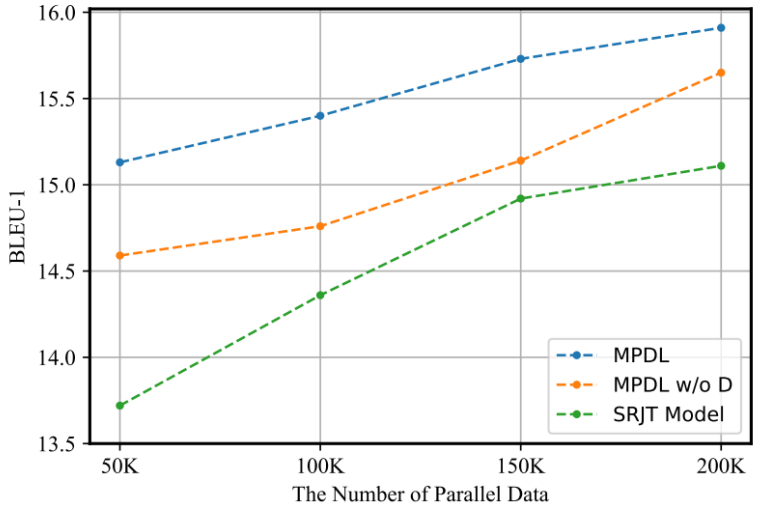
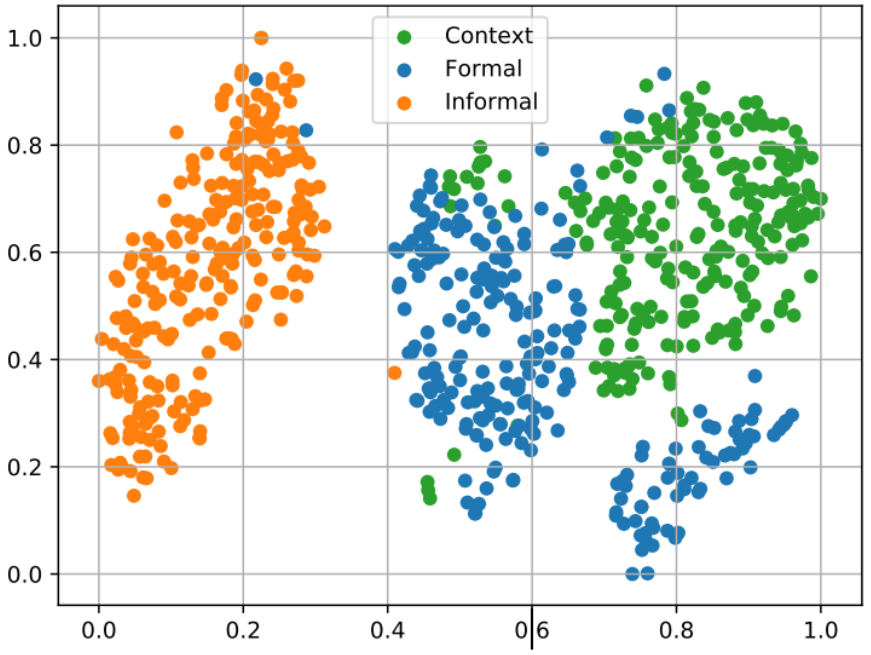


Figure 4: The impact of parallel conversation data quantity on model performance of BLEU-1.

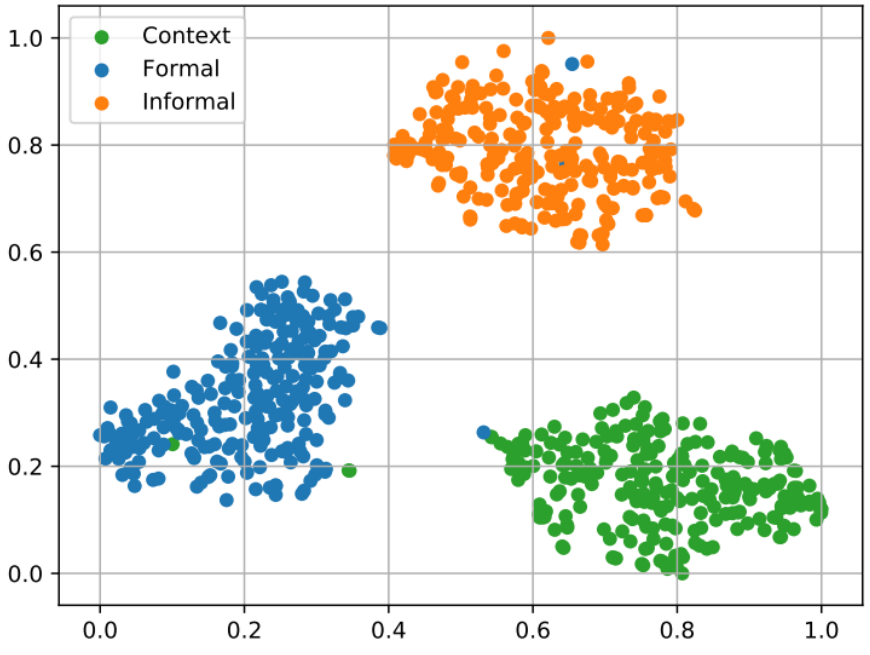
Model	BLEU-1	BLEU-2	Dist.	BERT	SVM
Ours	**12.0**	**1.38**	50.1	**84.4**	79.2
w/o Discri. D	10.9	1.33	**52.8**	83.1	79.4
w/o Multi-Pass	9.89	0.89	45.1	82.3	75.4
w/o DialoGPT	10.3	0.87	33.6	84.0	**79.7**

Table 3: Automatic evaluation average results of ablation models for responses in style formal and informal.

Experimentsal Results



(a) Without discriminators.



(b) With discriminators.

Figure 3: Visualizing the representations of the context, informal and formal text from TCFC dataset.

Experimentsal Results

Table 6: The performance of different discriminators in formal response on TCFC.

Discriminators	**BLEU-1**	**BLEU-2**	**Distinct**	**BERT**	**SVM**
3-layer feed-forward network	**16.5**	**2.07**	51.3	**98.6**	**97.1**
2-layer Transformer encoder	15.0	1.81	50.4	98.4	96.4
8-layer Transformer encoder	14.4	1.62	**52.8**	95.6	93.7

Table 7: The example of weights produced by discriminators.

Discriminators	Pseudo Text	Weight
Context D_x	She's such a great actress and I love her!	0.99
	What's the best way to get a job in my life?	0.87
	I don't know what to say.	0.19
Formality $D_{\tilde{y}}$	You have a very precious soul.	0.99
	I will share that pack with you!	0.68
	I would like to see U.	0.09
Shakespeare $D_{\tilde{y}}$	Tis a brave lady.	0.99
	I'll be here and you're welcome.	0.21
	I'm not sure what to do with it	0.14

Case Study

Table 4: The example responses produced by MPDL model and the baselines on TCFC dataset.

Context	A wedding ring? But why would you throw it away?
The target style is formal English (i.e., style S_1)	
Human	Are you sure? You must have surely at least pawned it.
MTask	I wish I had a chance to see that.
SFusion	They already let me set up! :)
SRJT	Yes, they should have given it to you. It is a wedding ring.
Ours	I have never thrown away a wedding ring before.
The target style is informal English (i.e., style S_0)	
Human	Right? Surely you would at least have pawned it.
S2S	I don't want to talk about it.
SFusion	I've emailed the proof when picks it in. I'll send it shortly.
SRJT	I'm sure you didn't throw it.
Ours	You know you wouldn't throw the engagement ring away.

Take-Away Message

- Preliminary solutions to some questions
- Low-resource knowledge selection for KGC
- Knowledge in reasoning
- Shared knowledge in
  - Multilinguality
  - Style

Q & A

Thank You!

Email: ruiyan@ruc.edu.cn