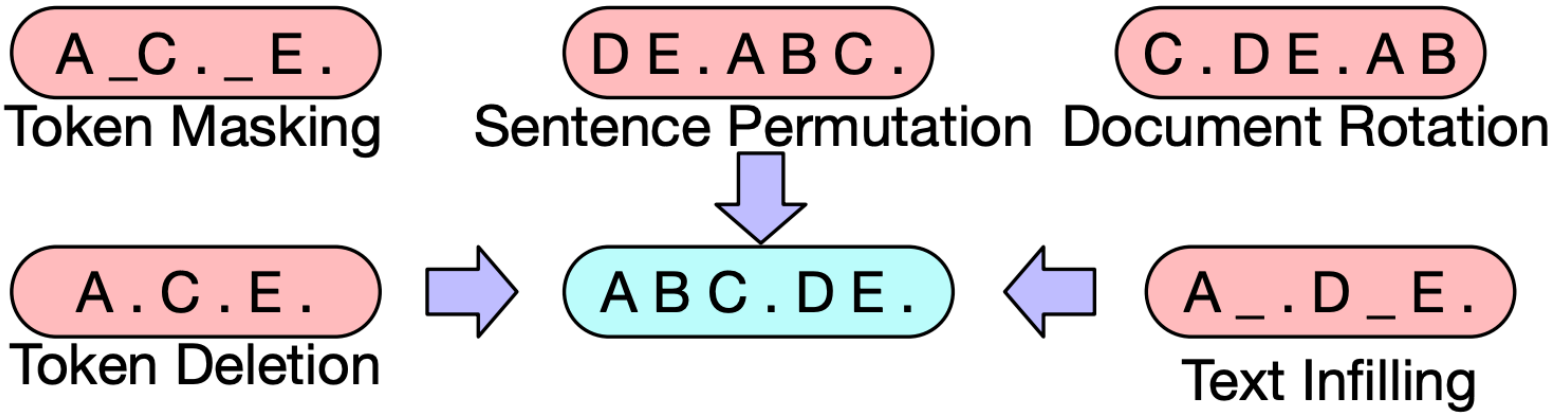
Self-supervised Tasks for Dialogue Context Modeling

Qi Jia

Self-Supervised Tasks

- BERT:
  - Token-level Masked LM
  - Next Sentence Prediction (NSP)
- BART:



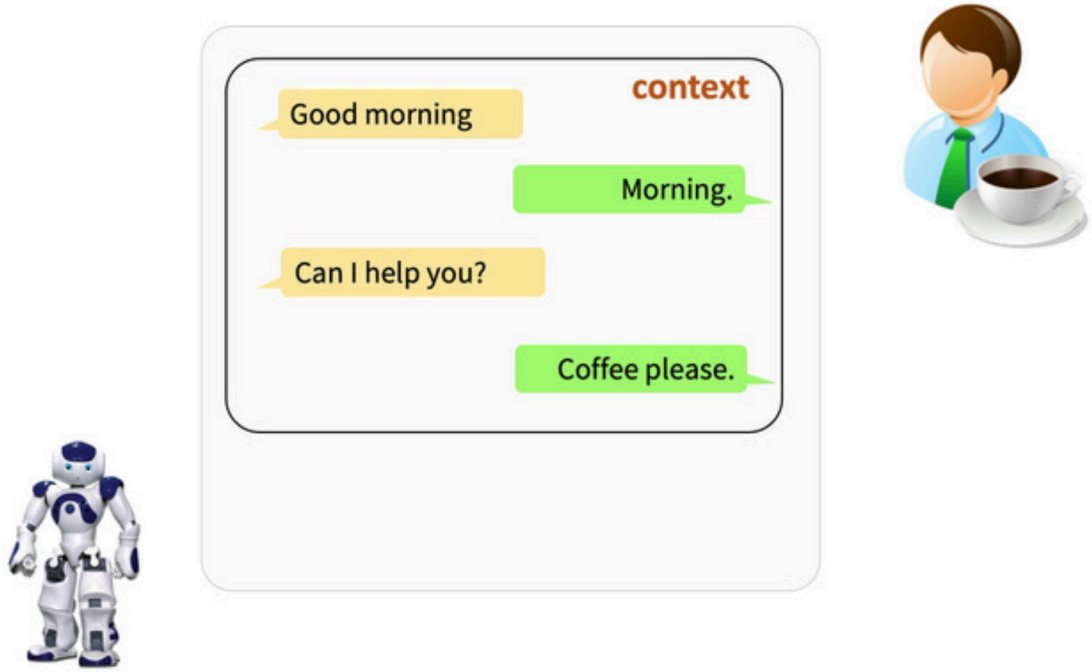
List

- DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances
- Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues

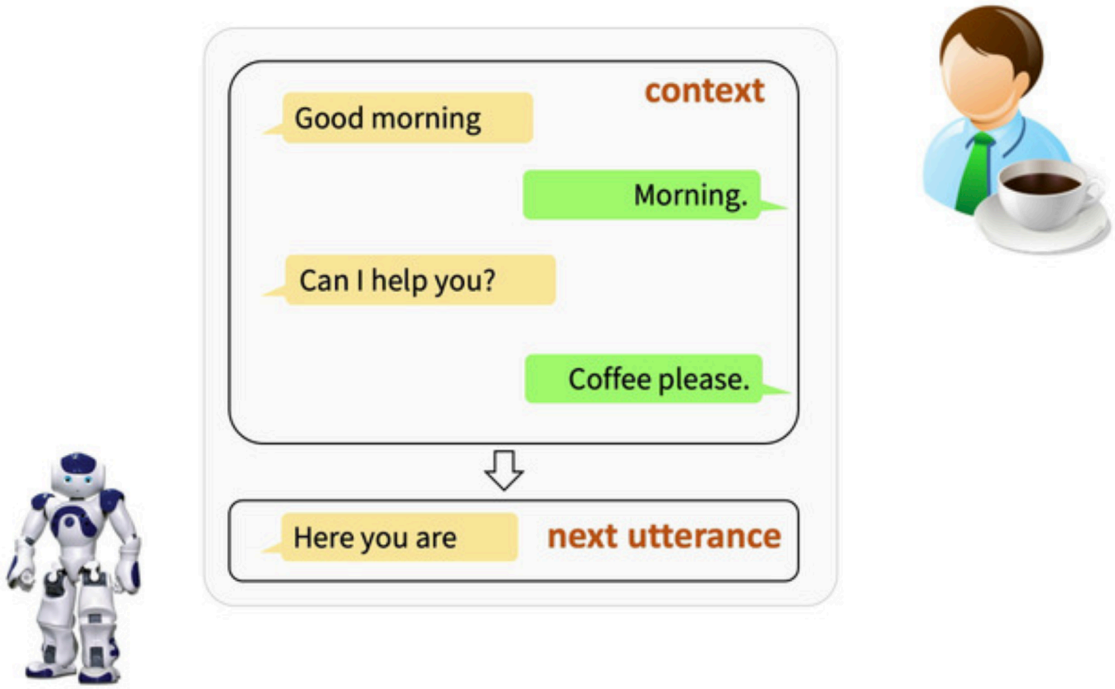
DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances

Xiaodong Gu, Kang Min Yoo, Jung-Woo Ha

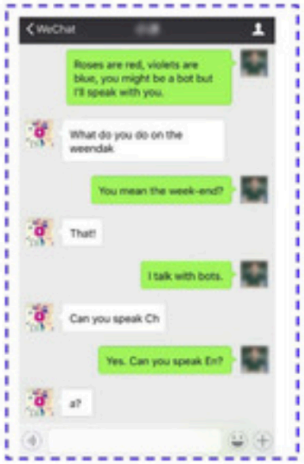
Open-Domain Dialogue Generation



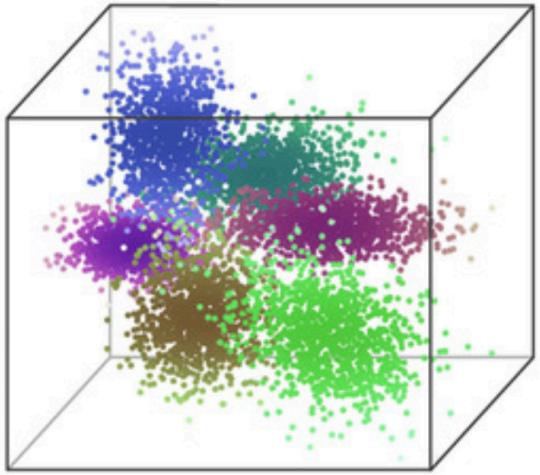
Open-Domain Dialogue Generation



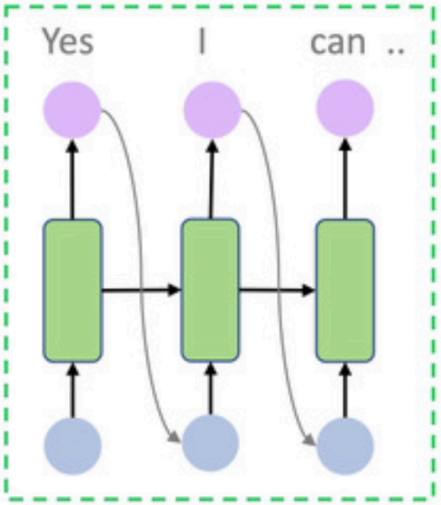
A General Recipe



Dialog Context

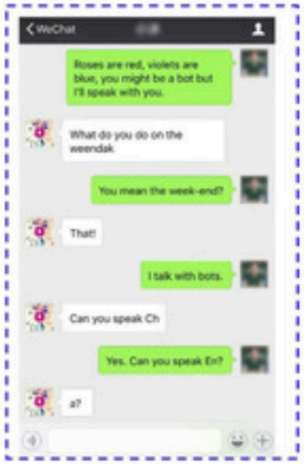


Dialog Context Representation

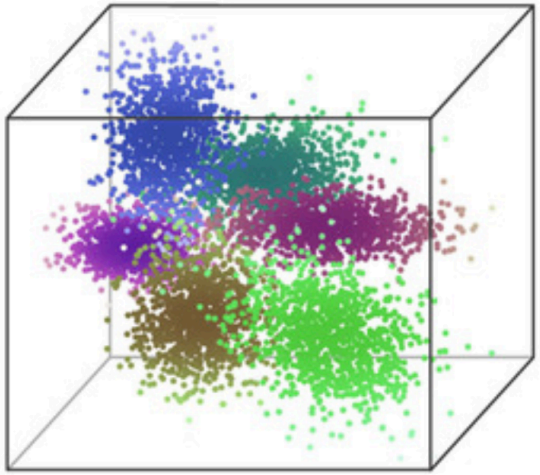


Response Generation

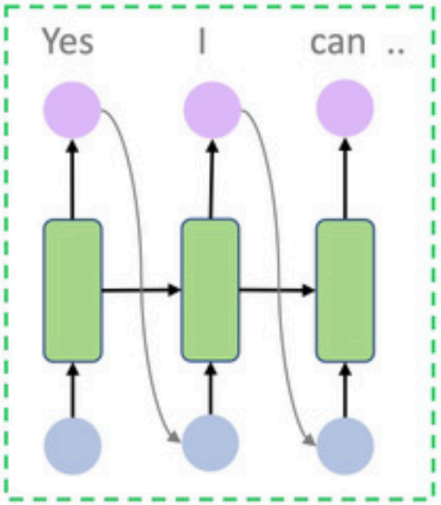
A General Recipe



Dialog Context



Dialog Context Representation

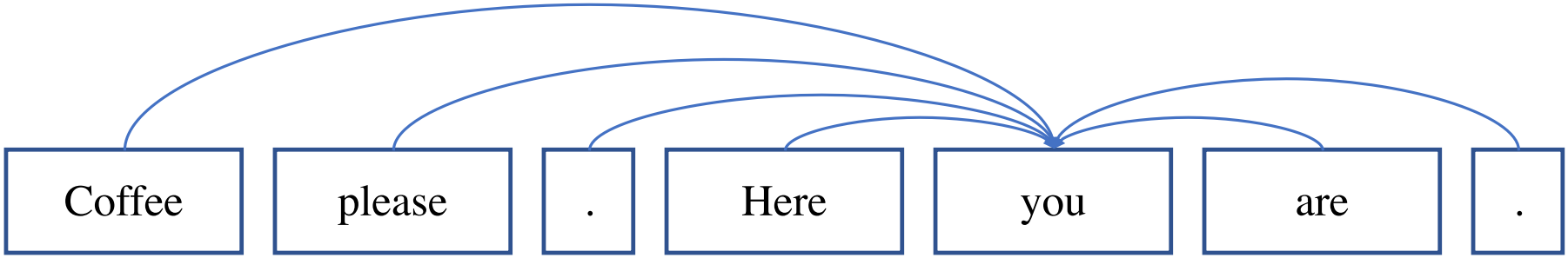


Response Generation

Learning the **rich representation** of dialogue contexts

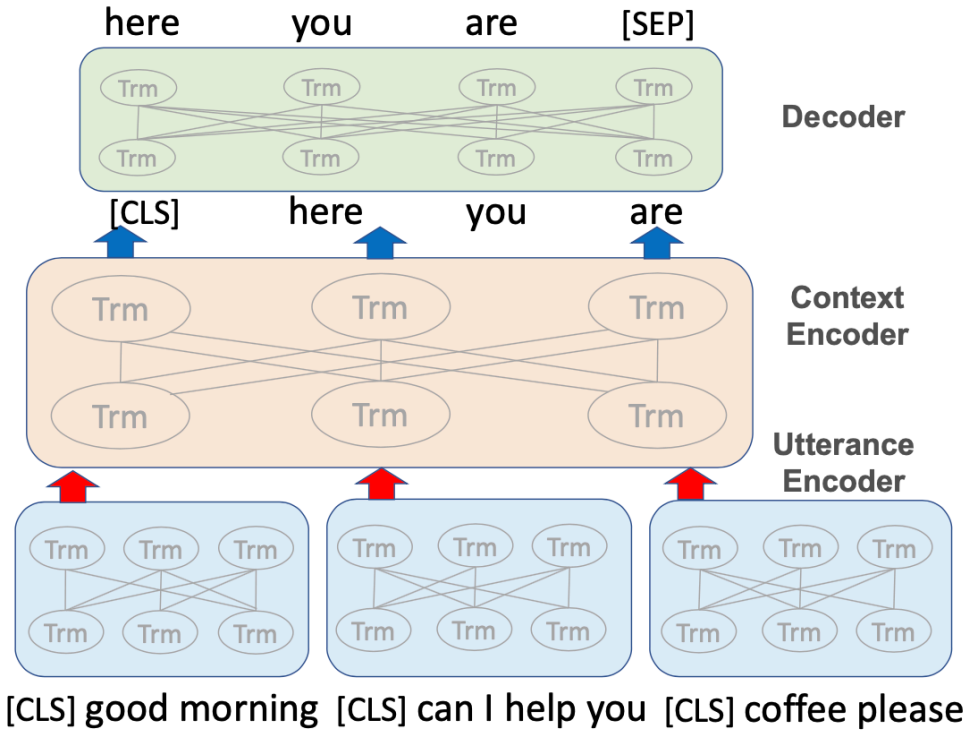
Limitations

- Encoding a **long** context.
- As a **linear sequence** of tokens.
- **Token-level** self-attention



DialogBERT

- **A hierarchical Transformer architecture**
  - **Utterance encoder** to encode utterances
  - **Context encoder** for discourse-level encoding of the entire context

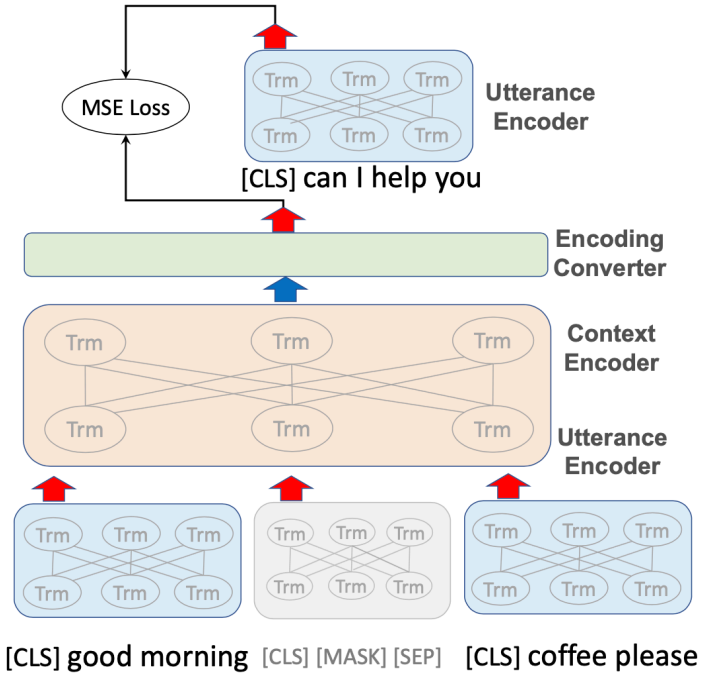


DialogueBERT

- **Two auxiliary training objectives**
  - Masked utterance regression (MURer)
  - Distributed utterance order ranking (DUOR)

Training: Masked Utterance Regression

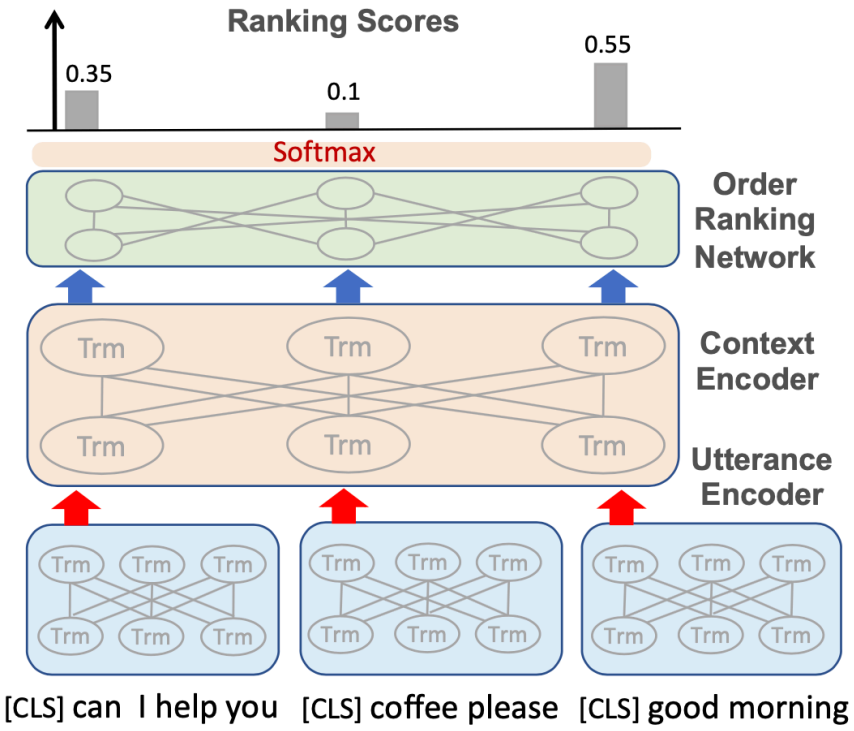
- Randomly **mask/replace** an utterance
- **Reconstruct** the vector of the original utterance through MSE loss



Training: Distributed utterance order ranking

- Organizing **randomly shuffled** utterances into a **coherent** context by minimizing KL-divergence

$$\hat{p} = (0.35, 0.1, 0.55)$$
$$p = softmax\left(\frac{2}{3}, \frac{3}{3}, \frac{1}{3}\right)$$



Experimental Setup - Datasets

- **Weibo** : large scale post-reply pairs from microblog
- **MultiWOZ** : human-to-human written conversations spanning over multiple domains
- **DailyDialog** : high-quality multi-turn chitchats designed for English learners.

Experimental Setup - Metrics

- **Perplexity (PPL)**
- **BLEU** : measures how many n-grams in a generated response overlap with those of the reference.
- **NIST** : a variant of BLEU that penalizes uninformative n-grams by assigning weights to n-grams according to their information gain.

Results – Automatic Evaluation

Model	Weibo			DailyDialog			MultiWOZ		
	PPL	**BLEU**	**NIST**	**PPL**	**BLEU**	**NIST**	**PPL**	**BLEU**	**NIST**
BART	34.15	6.5	6.6	22.92	7.59	9.78	5.28	11.83	19.69
DialoGPT	87.14	7.24	8.28	46.42	13.82	14.08	9.44	18.47	34.15
ContextPretrain-default	66.13	7.98	9.41	35.02	10.52	15.60	8.48	18.12	38.67
ContextPretrain-transformer	36.92	8.10	9.17	22.31	10.14	13.55	5.52	17.95	37.16
NUG only	31.73	8.25	9.45	22.16	14.41	22.33	5.06	19.50	49.83
NUG+MUR	28.08	8.49	10.12	21.14	13.81	21.50	4.96	19.89	52.74
NUG+DUOR	**27.29**	8.46	10.04	20.82	14.42	22.65	5.00	19.87	53.29
NUG+MUR+DUOR	27.30	**8.54**	**10.21**	**20.49**	**14.61**	**23.34**	**4.92**	**19.91**	**55.03**

Results – Human evaluations

- Randomly sampled 200 dialogues from DailyDialog.
- Generated responses from our model and other baselines.
- For each response pair, we ask three different annotators from AMT to evaluate the qualities and express preference.

Comparison	Coherence			Informativeness			Human-likeness		
	Win	**Tie**	**Loss**	**Win**	**Tie**	**Loss**	**Win**	**Tie**	**Loss**
Ours *vs.* BART	41%	34%	25%	38%	43%	19%	39%	45%	15%
Ours *vs.* DialoGPT	39%	45%	16%	40%	48%	12%	49%	43%	8%
Ours *vs.* ContextPretrain	44%	34%	22%	40%	42%	18%	39%	42%	19%

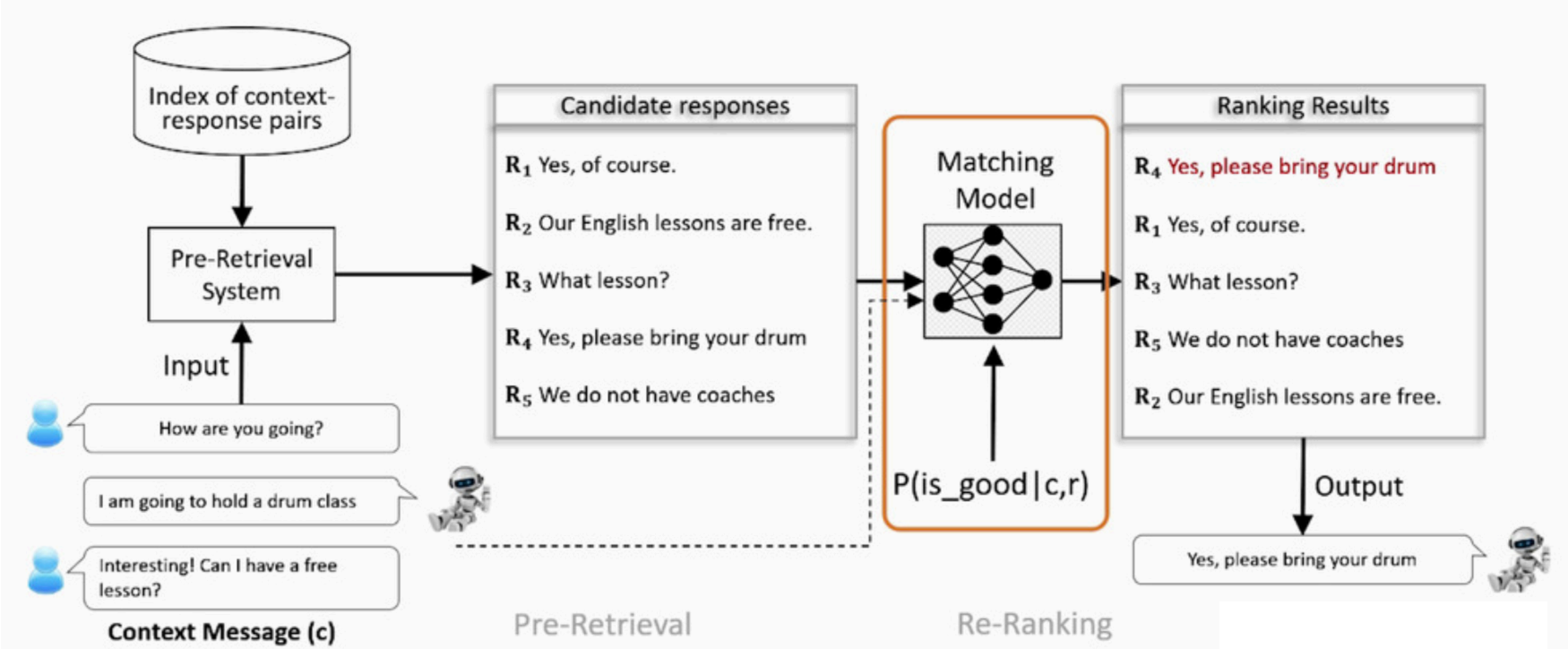
Conclusion

- DialogBERT -- a BERT extension for neural response generation
  - Hierarchical Transformer Encoder
  - Masked Utterance Regression
  - Distributed Utterance Order Ranking

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

Ruijian Xu, Chongyang Tao, et. al.

Retrieval-based Dialogue Systems



Context-Response Matching Models

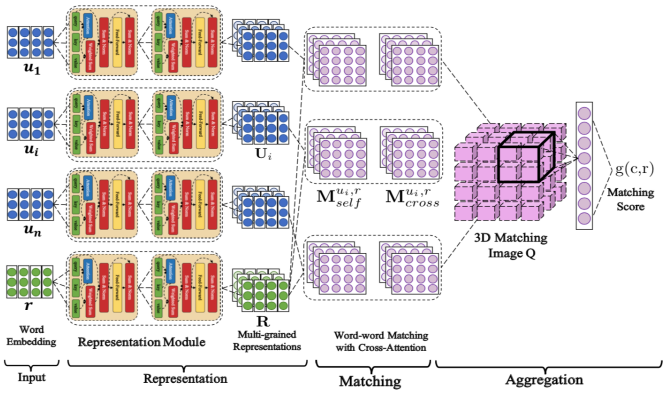
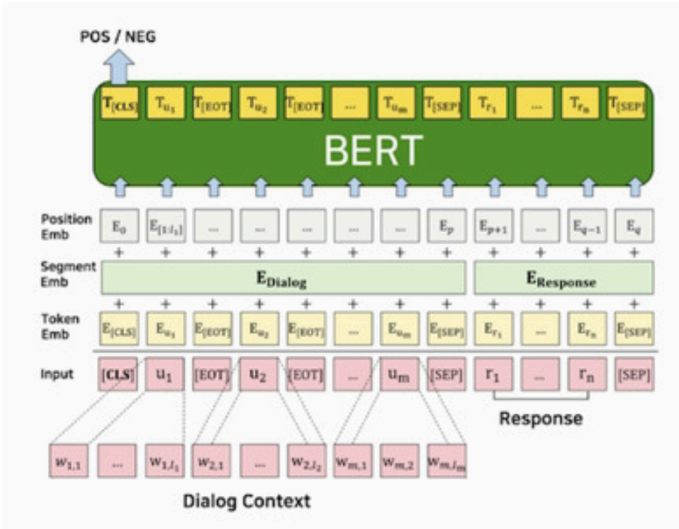


Figure 2: Overview of Deep Attention Matching Network.



Representation → Matching → Aggregation

Language Understanding

Semantic Matching

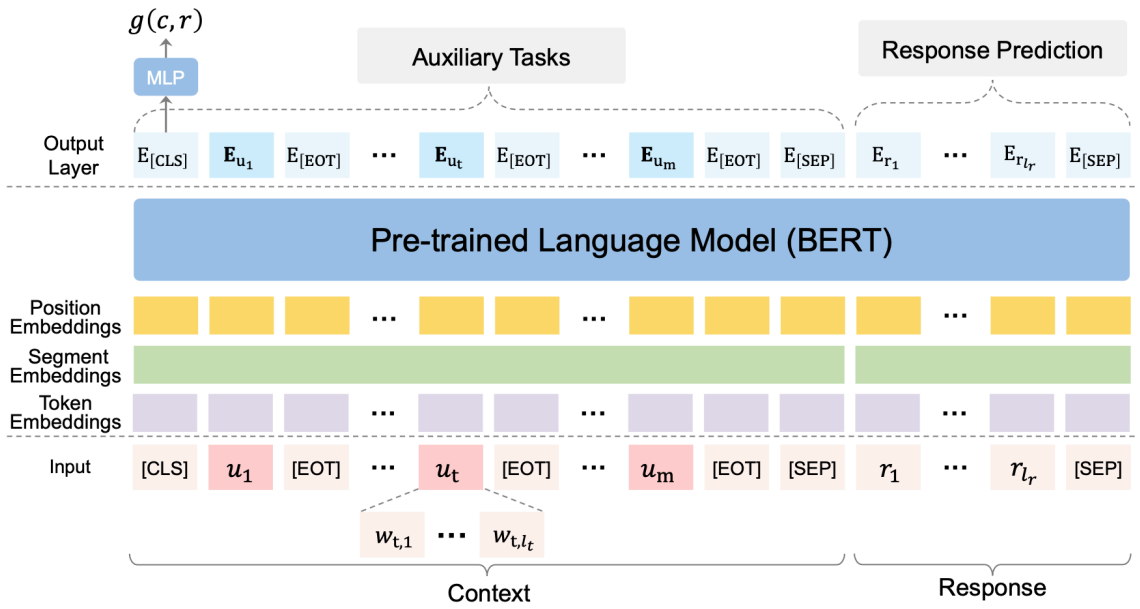
Multi-turn Aggregation

Motivations

- Building a context-response matching model with **various neural architectures** or pre-trained language models
- Learning with **a single response prediction task**
- Including **incoherence and inconsistency**

Methods

- **Self-supervised learning**
  - Construct various training signals with multi-turn dialogue
  - Jointly training with response matching task

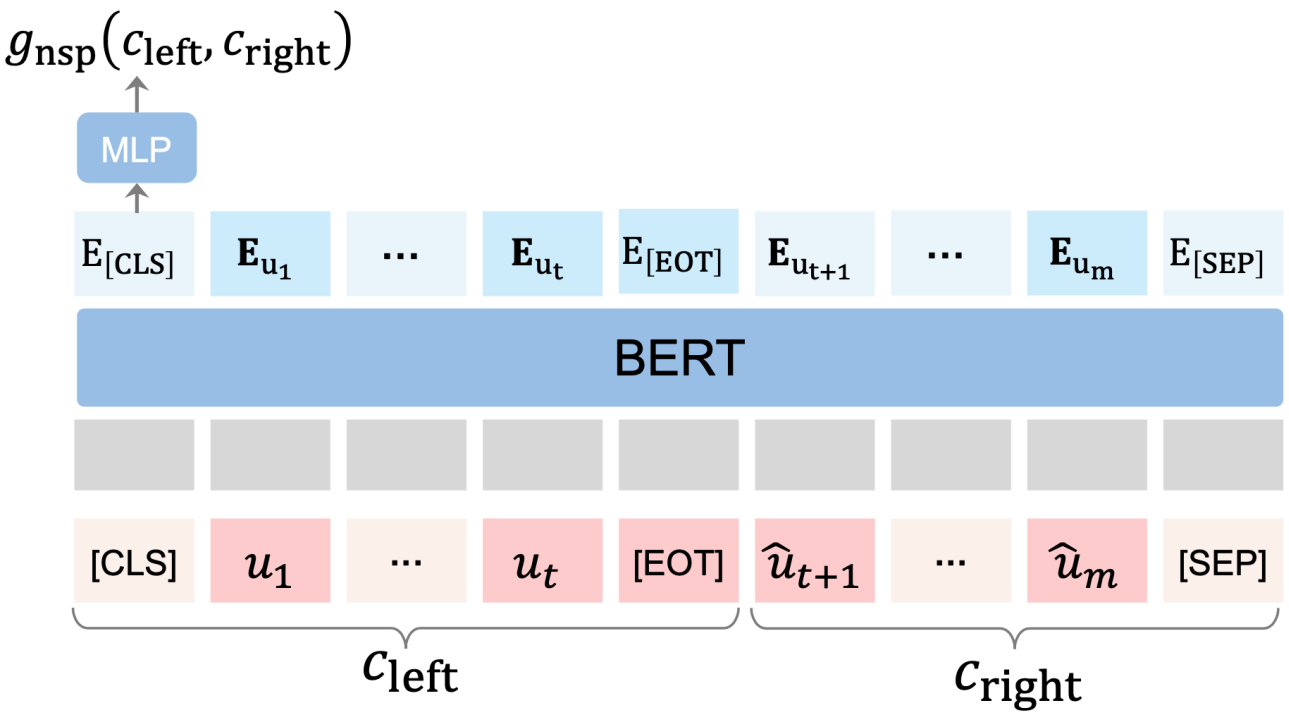


Better task-related representation
Better generalization ability

Self-supervised Tasks -1

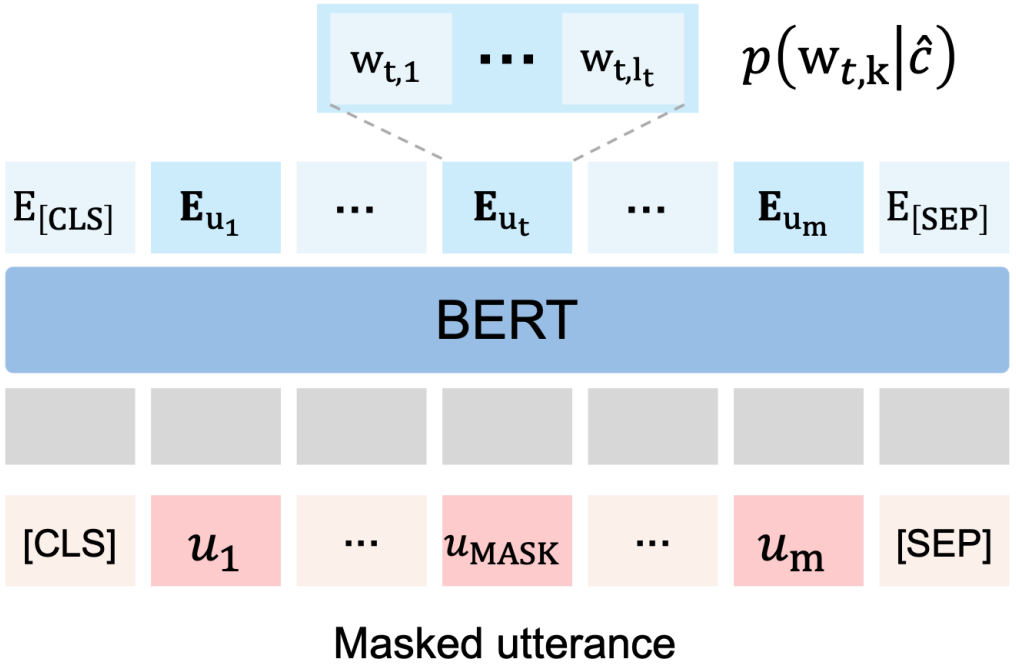
- **Next-session Prediction**
  - Split
  - Random replacement
  - Predict

Sequential Relationship



Self-supervised Tasks -2

- Utterance Restoration
  - mask all the tokens in an utterance
  - Predict a proper utterance

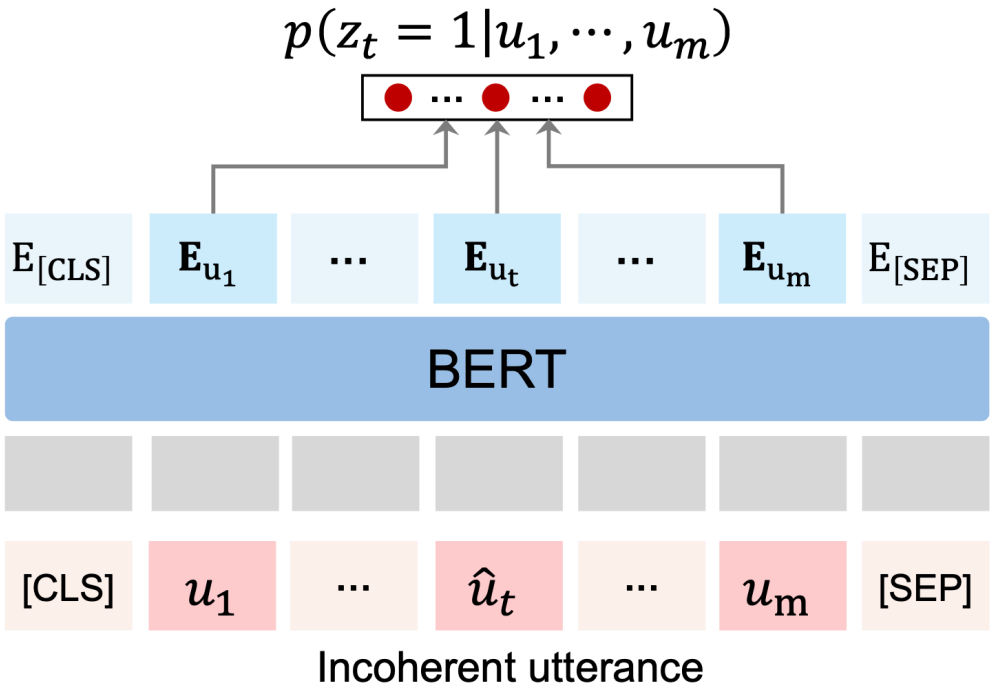


Semantic Connections

Self-supervised Tasks -3

- **Incoherence Detection**
  - Randomly select & replace
  - Recognize the incoherent utterance

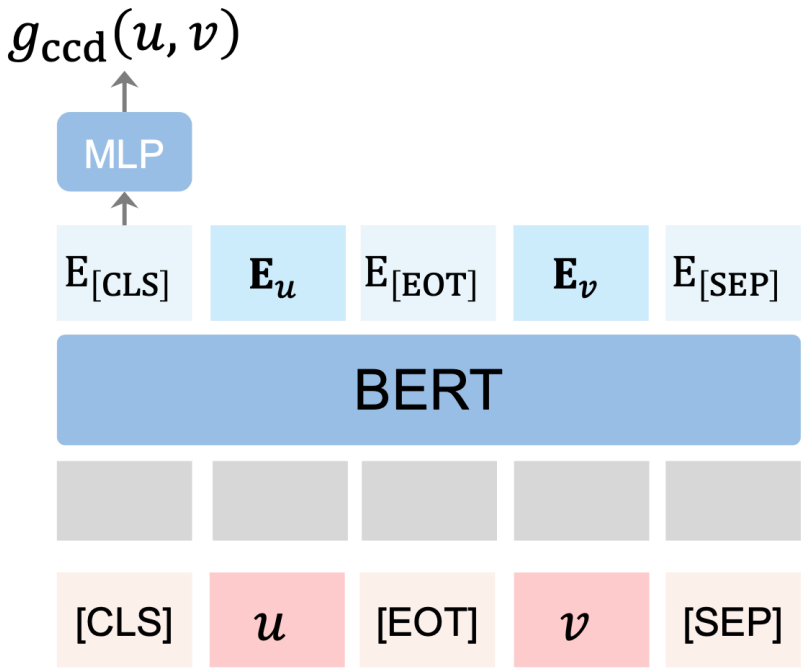
Sequential & Coherence



Self-supervised Tasks -4

- **Consistency Discrimination**
  - Sample u, v, v'
  - Hinge loss

Topic\Style Consistency



Experimental Setup - Datasets

- **Ubuntu Dialogue Corpus**
  - Multi-turn English dialogues about technical support
- **E-commerce Dialogue Corpus**
  - Real-world multi-turn dialogues between customers and customer service staff on Taobao.

Results - Automatic Evaluation

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., 2019b)	0.945	0.786	0.886	0.976	-	-	-
	IMN (Gu et al., 2019)	0.946	0.794	0.889	0.974	0.621	0.797	0.964
	ESIM (Chen & Wang, 2019)	0.950	0.796	0.874	0.975	0.570	0.767	0.948
	IoI (Tao et al., 2019a)	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)	0.954	0.817	0.904	0.977	0.610	0.814	0.973
	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

Results - Human evaluations

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

Conclusions

- Learning a context response matching model with multiple **auxiliary self-supervised** tasks
- The matching model can effectively learn **task-related knowledge** and produce better features for response selection
- The self-supervised tasks can bring **significant improvement** for various matching architectures

Take-aways

- Use self-supervised tasks to help especially with limited data
- The tasks should be designed considering the nature of your data