Paper Reading

Tian Lan

2021/7/1

Pretrained LM for Dialog Response Selection

- **Task Formulation of the Dialog Response Selection**
- **BERT-VFT** An Effective Domain Adaptive Post-Training Method for BERT in Response Selection (Interspeech 2020)
- **SA-BERT** Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots (CIKM 2020)
- **UMS-BERT** Do Response Selection Models Really Know What's Next? Utterance Manipulation Strategies for Multi-turn Response Selection (AAAI 2021)
- **BERT-SL** Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues (AAAI 2021)
- **HCL** Dialogue Response Selection with Hierarchical Curriculum Learning (ACL 2021)
- **BERT-FP** Fine-grained Post-training for Improving Retrieval-based Dialogue Systems (NAACL 2021)

Task Formulation

- **Input**: multi-turn conversation context c, and one candidate r
- **Output**: the matching degree $s = f(c, r)$, where f is the model

  f is the BERT or the RNN model
- **Benchmarks**: Ubuntu-v1, Douban, E-Commerce

  1 million samples for training (pos:neg = 1:1), 1000 sessions for testing (pos:neg=1:9)
- **Evaluation Metric:** Information retrieval matric (recall)
  - $R_2@1$, $R_{10}@1$, $R_{10}@2$, $R_{10}@5$, MRR, MAP

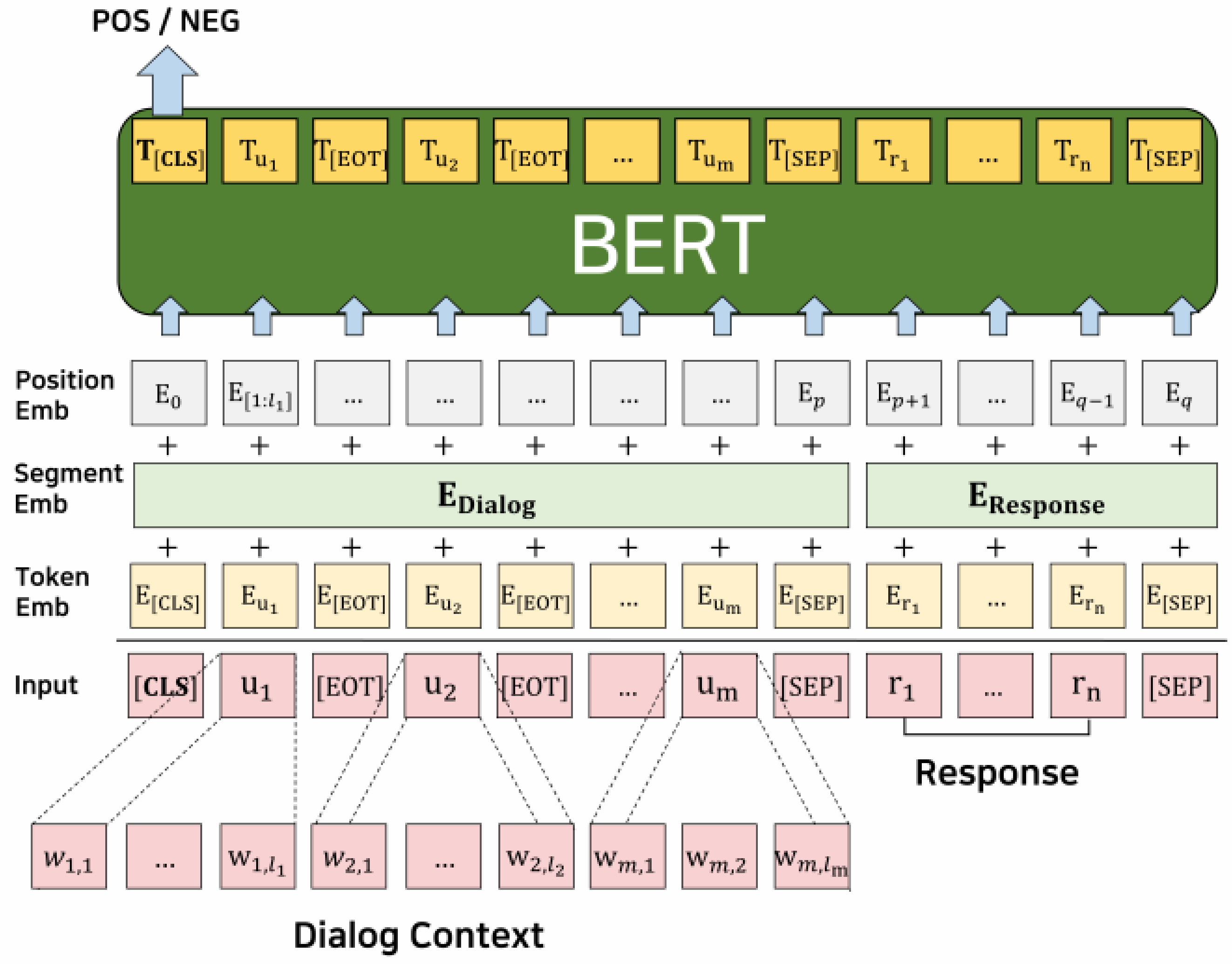
BERT-VFT

An Effective Domain Adaptive Post-Training Method for BERT in Response Selection

Taesun Whang[1*] *Dongyub Lee*[2] *Chanhee Lee*[3] *Kisu Yang*[3] *Dongsuk Oh*[3] *Heuiseok Lim*[3]

BERT-VFT

- **The first work to utilize the BERT** for dialog response selection task, and achieve the SOTA performance
- Use **post-train** to improve the performance of BERT further



BERT-VFT experiment

- The results on Ubuntu-v1 corpus prove the effectiveness of the **BERT** model for this task
- The post-train (**BERT-VFT**) brings very huge improvement

Model	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
DualEncoder$_{rnn}$	0.403	0.547	0.819
DualEncoder$_{cnn}$	0.549	0.684	0.896
DualEncoder$_{lstm}$	0.638	0.784	0.949
DualEncoder$_{bilstm}$	0.630	0.780	0.944
MultiView	0.662	0.801	0.951
SMN	0.726	0.847	0.961
AK-DE-biGRU	0.747	0.868	0.972
DUA	0.752	0.868	0.962
DAM	0.767	0.874	0.969
MRFN	0.786	0.886	0.976
IoI	0.796	0.894	0.974
MSN	<u>0.800</u>	<u>0.899</u>	<u>0.978</u>
BERT$_{base}$	0.817	0.904	0.977
BERT-DPT	0.851	0.924	0.984
BERT-VFT	**0.855**	**0.928**	**0.985**
BERT-VFT(DA)	**0.858**	**0.931**	**0.985**

Table 2: *Model comparison on Ubuntu Corpus V1.*

BERT-VFT conclusion

- Advantage:
  - The first work to introduce the BERT model into dialog response selection task
  - Test the effectiveness of the post-train for this downstream task
- Disadvantage:
  - Only test on Ubuntu-v1 corpus, missing the experiments on other benchmarks, such as Douban, E-commerce
  - Missing the experiments and analyse of the dual-encoder(BERT) model

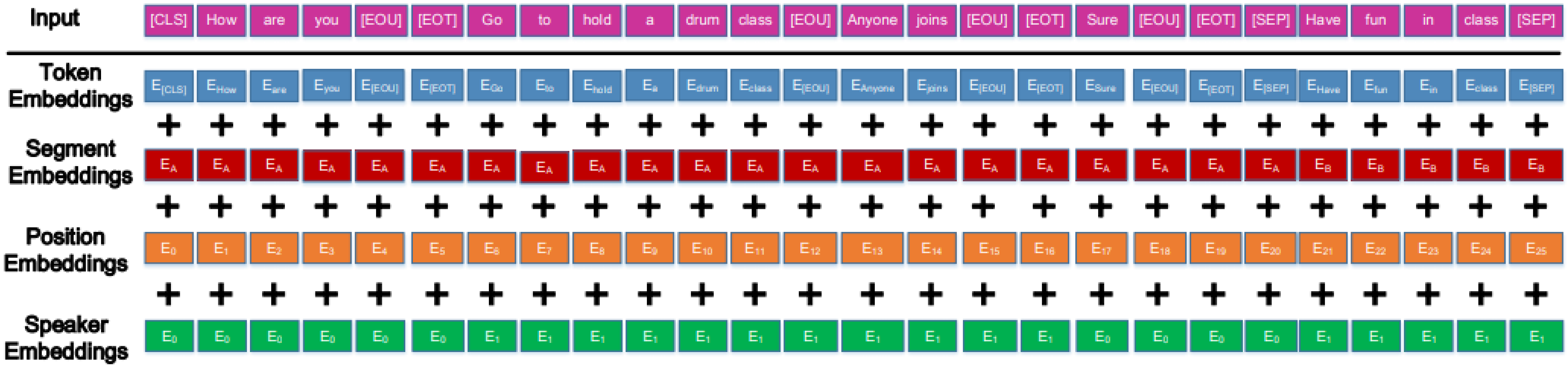
SA-BERT

Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots

Jia-Chen Gu[1], Tianda Li[2], Quan Liu[1,3], Zhen-Hua Ling[1], Zhiming Su[3], Si Wei[3], Xiaodan Zhu[2]
[1]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
[2]ECE & Ingenuity Labs, Queen's University, Kingston, Canada
[3]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research, Hefei, China
gujc@mail.ustc.edu.cn,tianda.li/xiaodan.zhu@queensu.ca,quanliu/zhling@ustc.edu.cn,zmsu/siwei@iflytek.com

SA-BERT

- Leverage the **speaker information** into the multi-turn dialog conversation
- Rich experiments on 5 datasets
- Post-train is used



SA-BERT experiment

Table 2: Evaluation results of SA-BERT and previous methods on the Ubuntu Dialogue Corpus V1 and V2.

	Ubuntu Corpus V1				Ubuntu Corpus V2			
	$R_2@1$	**$R_{10}@1$**	**$R_{10}@2$**	**$R_{10}@5$**	**$R_2@1$**	**$R_{10}@1$**	**$R_{10}@2$**	**$R_{10}@5$**
SMN [15]	0.926	0.726	0.847	0.961	-	-	-	-
DUA [18]	-	0.752	0.868	0.962	-	-	-	-
DAM [20]	0.938	0.767	0.874	0.969	-	-	-	-
MRFN [12]	0.945	0.786	0.886	0.976	-	-	-	-
IMN [4]	0.946	0.794	0.889	0.974	0.945	0.771	0.886	0.979
IoI [13]	0.947	0.796	0.894	0.974	-	-	-	-
MSN [17]	-	0.800	0.899	0.978	-	-	-	-
BERT	0.950	0.808	0.897	0.975	0.950	0.781	0.890	0.980
SA-BERT	**0.965**	**0.855**	**0.928**	**0.983**	**0.963**	**0.830**	**0.919**	**0.985**

Table 3: Evaluation results of SA-BERT and previous methods on the Douban Corpus and E-commerce Corpus.

	Douban Conversation Corpus						E-commerce Corpus		
	MAP	**MRR**	**P@1**	**$R_{10}@1$**	**$R_{10}@2$**	**$R_{10}@5$**	**$R_{10}@1$**	**$R_{10}@2$**	**$R_{10}@5$**
SMN [15]	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DUA [18]	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
DAM [20]	0.550	0.601	0.427	0.254	0.410	0.757	-	-	-
MRFN [12]	0.571	0.617	0.448	0.276	0.435	0.783	-	-	-
IMN [4]	0.570	0.615	0.433	0.262	0.452	0.789	0.621	0.797	0.964
IoI [13]	0.573	0.621	0.444	0.269	0.451	0.786	0.563	0.768	0.950
MSN [17]	0.587	0.632	0.470	0.295	0.452	0.788	0.606	0.770	0.937
BERT	0.591	0.633	0.454	0.280	0.470	0.828	0.610	0.814	0.973
SA-BERT	**0.619**	**0.659**	**0.496**	**0.313**	**0.481**	**0.847**	**0.704**	**0.879**	**0.985**

SA-BERT conclusion

- Advantage
  - It is reasonable to use the speaker information in the multi-turn conversation
- Disadvantage
  - Missing the **ablation study of the post-train** procedure. It is unclear whether the improvement is made by post-train or the speaker information in their experiment.

UMS-BERT


Do Response Selection Models Really Know What's Next?
Utterance Manipulation Strategies For Multi-turn Response Selection
Taesun Whang1* Dongyub Lee2* Dongsuk Oh3 Chanhee Lee3
Kijong Han4 Dong-hun Lee4 Saebyeok Lee1,3†
1Wisenut Inc.
2Kakao Corp.
3Korea University
4Kakao Enterprise Corp.

UMS-BERT

- The response selection task alone is insufficient. In this work, three well designed **auxiliary tasks** are used.
  - Insertion

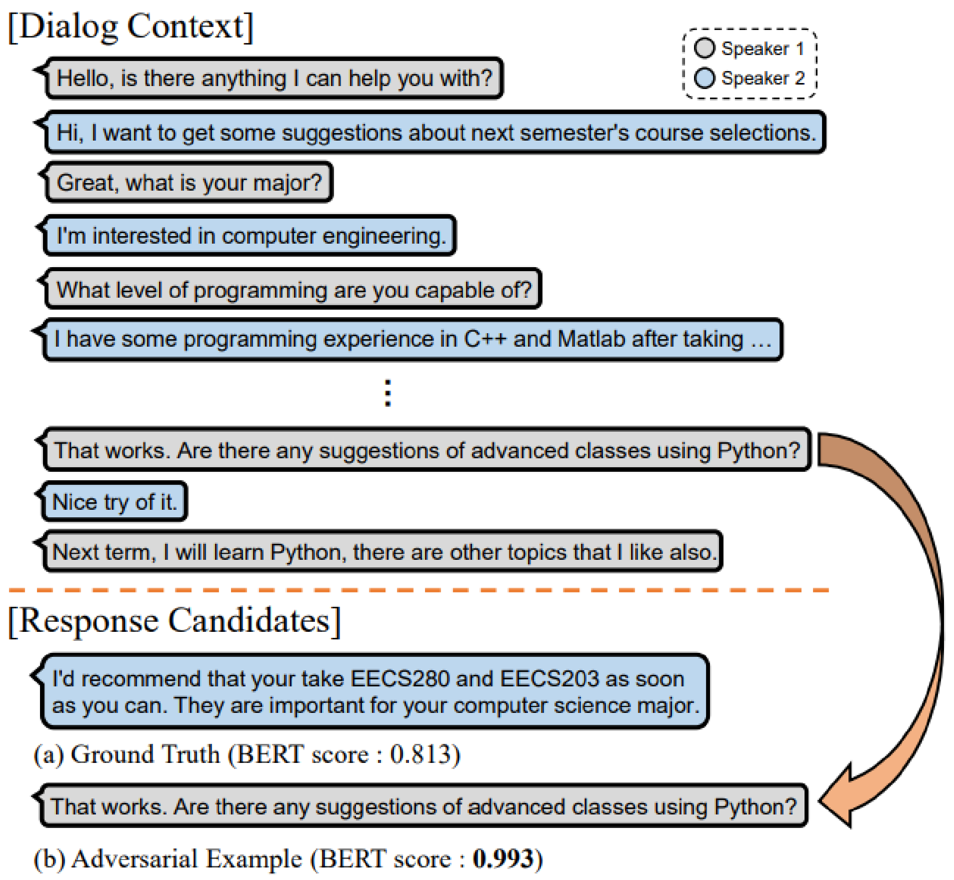
$$\mathbf{X}_{\text{INS}} = [[\text{CLS}]\ [\text{INS}]_1\, u_1 [\text{INS}]_2\, u_2 \ldots u_{t-1}\ [\text{INS}]_t\, u_{t+1} \ldots u_k\ [\text{INS}]_k\ [\text{SEP}]\, u_t\ [\text{SEP}]]$$

  - Deletion

$$\mathbf{X}_{\text{DEL}} = [[\text{CLS}]\ [\text{DEL}]_1\, u_1\ [\text{DEL}]_2\, u_2 \ldots [\text{DEL}]_t\ u^{rand}\ [\text{DEL}]_{t+1}\, u_t \ldots [\text{DEL}]_{k+1}\, u_k\ [\text{SEP}]]$$

  - Search

$$\mathbf{X}_{\text{SRCH}} = [[\text{CLS}]\ [\text{SRCH}]_1\, u'_1 [\text{SRCH}]_2\, u'_2 \ldots [\text{SRCH}]_t\, u'_t \ldots u'_{k-1} [\text{SEP}]\ u_k\ [\text{SEP}]]$$



UMS-BERT experiment

Models	Ubuntu			Douban						E-commerce		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	P@1	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
CNN (Kadlec, Schmid, and Kleindienst 2015)	0.549	0.684	0.896	0.417	0.440	0.226	0.121	0.252	0.647	0.328	0.515	0.792
LSTM (Kadlec, Schmid, and Kleindienst 2015)	0.638	0.784	0.949	0.485	0.537	0.320	0.187	0.343	0.720	0.365	0.536	0.828
BiLSTM (Kadlec, Schmid, and Kleindienst 2015)	0.630	0.780	0.944	0.479	0.514	0.313	0.184	0.330	0.716	0.365	0.536	0.825
MV-LSTM (Wan et al. 2016)	0.653	0.804	0.946	0.498	0.538	0.348	0.202	0.351	0.710	0.412	0.591	0.857
Match-LSTM(Wang and Jiang 2016)	0.653	0.799	0.944	0.500	0.537	0.345	0.202	0.348	0.720	0.410	0.590	0.858
Multi-View (Zhou et al. 2016)	0.662	0.801	0.951	0.505	0.543	0.342	0.202	0.350	0.729	0.421	0.601	0.861
DL2R (Yan, Song, and Wu 2016)	0.626	0.783	0.944	0.488	0.527	0.330	0.193	0.342	0.705	0.399	0.571	0.842
SMN (Wu et al. 2017)	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DUA (Zhang et al. 2018)	0.752	0.868	0.962	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
DAM (Zhou et al. 2018)	0.767	0.874	0.969	0.550	0.601	0.427	0.254	0.410	0.757	0.526	0.727	0.933
IoI (Tao et al. 2019b)	0.796	0.894	0.974	0.573	0.621	0.444	0.269	0.451	0.786	0.563	0.768	0.950
MSN (Yuan et al. 2019)	0.800	0.899	0.978	0.587	0.632	0.470	0.295	0.452	0.788	0.606	0.770	0.937
BERT (Gu et al. 2020)	0.808	0.897	0.975	0.591	0.633	0.454	0.280	0.470	0.828	0.610	0.814	0.973
BERT-SS-DA (Lu et al. 2020)	0.813	0.901	0.977	0.602	0.643	0.458	0.280	0.491	0.843	0.648	0.843	0.980
SA-BERT (Gu et al. 2020)	0.855	0.928	0.983	0.619	0.659	0.496	0.313	0.481	0.847	0.704	0.879	0.985
BERT (ours)	0.820	0.906	0.978	0.597	0.634	0.448	0.279	0.489	0.823	0.641	0.824	0.973
ELECTRA	0.826	0.908	0.978	0.602	0.642	0.465	0.287	0.483	0.839	0.609	0.804	0.965
UMS_{BERT}	0.843	0.920	0.982	0.597	0.639	0.466	0.285	0.471	0.829	0.674	0.861	0.980
$\text{UMS}_{\text{ELECTRA}}$	0.854	0.929	0.984	0.608	0.650	0.472	0.291	0.488	0.845	0.648	0.831	0.974
BERT+	0.862	0.935	0.987	0.609	0.645	0.463	0.290	**0.505**	0.838	0.725	0.890	0.984
ELECTRA+	0.861	0.932	0.985	0.612	0.655	0.480	0.301	0.499	0.836	0.673	0.835	0.974
$\text{UMS}_{\text{BERT+}}$	**0.875**†	**0.942**†	**0.988**†	**0.625**	**0.664**	**0.499**	**0.318**	0.482	**0.858**	**0.762**	**0.905**	**0.986**
$\text{UMS}_{\text{ELECTRA+}}$	**0.875**	0.941	**0.988**	0.623	0.663	0.492	0.307	0.501	0.851	0.707	0.853	0.974

UMS-BERT experiment

	Auxiliary Tasks	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MRR
1	None	0.826	0.908	0.978	0.890
2	INS	0.836	0.917	0.980	0.897
3	DEL	0.848	0.924	0.983	0.905
4	SRCH	0.834	0.915	0.981	0.896
5	INS + DEL	0.853	0.927	0.984	0.909
6	INS + SRCH	0.841	0.920	0.982	0.901
7	DEL + SRCH	0.852	0.927	0.983	0.908
8	INS + DEL + SRCH	**0.854**	**0.929**	**0.984**	**0.910**

- Each strategy contributes to the performance
- Contributions order: DEL > INS ≈ SRCH

Approach	Model	Original $R_{10}@1$	Original MRR	Adversarial $R_{10}@1$	Adversarial MRR
Baselines	BERT	0.820	0.887	0.199	0.561
	BERT+	**0.862**	**0.915**	0.203	0.573
	ELECTRA	0.826	0.890	0.304	0.614
	ELECTRA+	0.861	0.914	**0.329**	**0.636**
	Avg	0.842	0.902	0.259	0.596
UMS	BERT	0.843	0.902	0.310	0.622
	BERT+	**0.875**	**0.923**	0.363	0.656
	ELECTRA	0.854	0.910	0.397	0.668
	ELECTRA+	**0.875**	0.922	**0.437**	**0.692**
	Avg	0.862	0.914	0.377	0.660

- Adversarial candidates are used to examine the robustness.
- Adversarial candidate are randomly sampled from the multi-turn conversation context.
- UMS is more robust than BERT

UMS-BERT conclusion

- Advantage
  - Auxiliary tasks is straightforward and reasonable, which is very similar to the BERT pre-training procedure (NSP, MLM, SOP, …)
- Disadvantage
  - After reading the codes of their codes, I find that they create the negative samples for each strategy, and **the size of the training dataset are 3x larger than the previous training protocol**. More experiments should be added to prove the improvements are brought from the strategy other than the more negative samples.

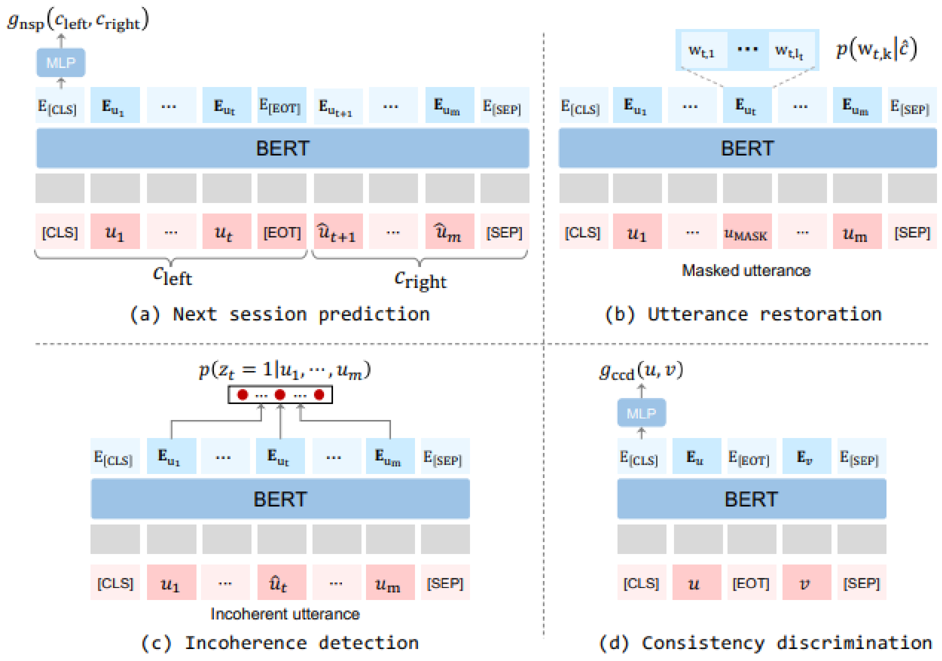
BERT-SL

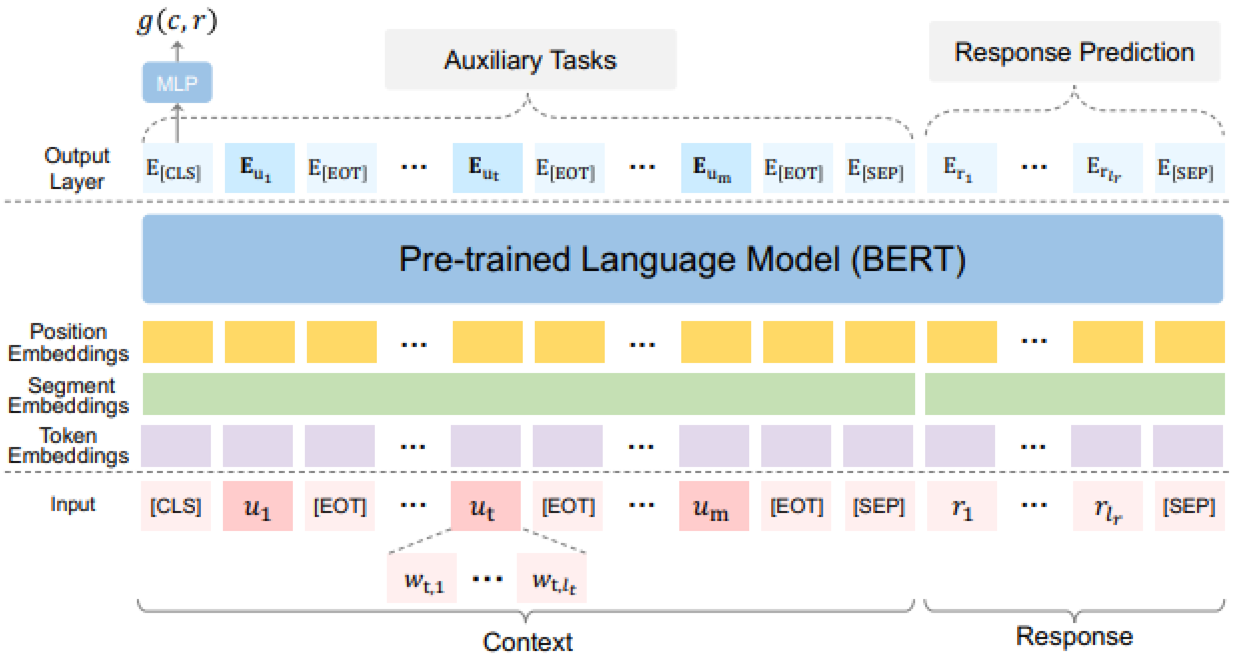
Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues

Ruijian Xu [*1] Chongyang Tao [*2] Daxin Jiang [2] Xueliang Zhao [3] Dongyan Zhao [1] Rui Yan [1]

BERT-SL

- Very similar to UMS-BERT.
- Four auxiliary tasks are used





BERT-SL experiment

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

- BERT-SL achieves the SOTA performance
- Ablation study prove the effectiveness of each strategy

BERT-SL conclusion

Same as the UMS-BERT

HCL

Dialogue Response Selection with Hierarchical Curriculum Learning

Yixuan Su[♠,*] Deng Cai[♡] Qingyu Zhou[◇] Zibo Lin[◇] Simon Baker[♠]
Yunbo Cao[◇] Shuming Shi[◇] Nigel Collier[♠] Yan Wang[◇]
[♠]Language Technology Lab, University of Cambridge
[♡]The Chinese University of Hong Kong
[◇]Tencent Inc.

HCL

- Leverage the curriculum learning to train the model in **easy-to-difficult schema**
- Hierarchical curriculum learning are proposed
  - Corpus-level:

    The ranking model (**fast dual encoder ranking model**) are used to measure the **difficulty** of each (c, r) pair. Easy pairs are first used for training, then the hard pairs.
  - Instance-level:

    The ranking model are used to measure the **difficulty** of negative samples for the context c. Easy negative samples are first used for training, then the hard pairs.

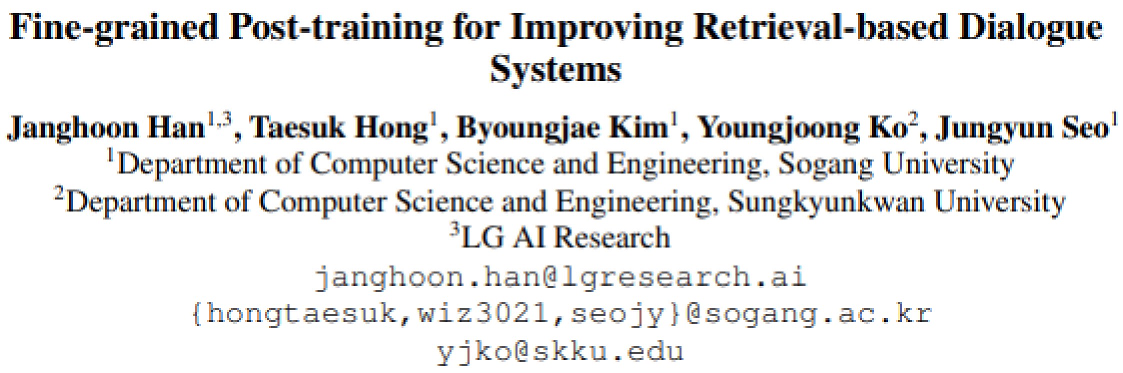
HCL experiment

Model	Douban						Ubuntu				E-Commerce		
	MAP	MRR	P@1	R_{10}@1	R_{10}@2	R_{10}@5	R_2@1	R_{10}@1	R_{10}@2	R_{10}@5	R_{10}@1	R_{10}@2	R_{10}@5
RNN	0.390	0.422	0.208	0.118	0.223	0.589	0.768	0.403	0.547	0.819	0.325	0.463	0.775
CNN	0.417	0.440	0.226	0.121	0.252	0.647	0.848	0.549	0.684	0.896	0.328	0.515	0.792
LSTM	0.485	0.527	0.320	0.187	0.343	0.720	0.901	0.638	0.784	0.949	0.365	0.536	0.828
BiLSTM	0.479	0.514	0.313	0.184	0.330	0.716	0.895	0.630	0.780	0.944	0.355	0.525	0.825
MV-LSTM	0.498	0.538	0.348	0.202	0.351	0.710	0.906	0.653	0.804	0.946	0.412	0.591	0.857
Match-LSTM	0.500	0.537	0.345	0.202	0.348	0.720	0.904	0.653	0.799	0.944	0.410	0.590	0.858
DL2R	0.488	0.527	0.330	0.193	0.342	0.705	0.899	0.626	0.783	0.944	0.399	0.571	0.842
Multi-View	0.505	0.543	0.342	0.202	0.350	0.729	0.908	0.662	0.801	0.951	0.421	0.601	0.861
DUA	0.551	0.599	0.421	0.243	0.421	0.780	-	0.752	0.868	0.962	0.501	0.700	0.921
DAM	0.550	0.601	0.427	0.254	0.410	0.757	0.938	0.767	0.874	0.969	0.526	0.727	0.933
MRFN	0.571	0.617	0.448	0.276	0.435	0.783	0.945	0.786	0.886	0.976	-	-	-
IOI	0.573	0.621	0.444	0.269	0.451	0.786	0.947	0.796	0.894	0.974	0.563	0.768	0.950
SMN	0.529	0.569	0.397	0.233	0.396	0.724	0.926	0.726	0.847	0.961	0.453	0.654	0.886
MSN	0.587	0.632	0.470	0.295	0.452	0.788	-	0.800	0.899	0.978	0.606	0.770	0.937
SA-BERT	0.619	0.659	0.496	0.313	0.481	0.847	0.965	0.855	0.928	0.983	0.704	0.879	0.985
SMN+HCL	0.575	0.620	0.446	0.281	0.452	0.807	0.947	0.777	0.885	0.981	0.507	0.723	0.935
MSN+HCL	0.620	0.668	0.507	0.321	0.508	0.841	0.969	0.826	0.924	0.989	0.642	0.814	0.968
SA-BERT+HCL	**0.639**	**0.681**	**0.514**	**0.330**	**0.531**	**0.858**	**0.977**	**0.867**	**0.940**	**0.992**	**0.721**	**0.896**	**0.993**

HCL experiment

- Advantage
  - Rich experiments:
    - Traditional evaluation protocol
    - Different learning strategy
    - Different learning architecture (RNN, Transformers, BERT)
    - Ablation study
- Disadvantage
  - Hard to implement
  - Lots of hyper-parameters during training

BERT-FP

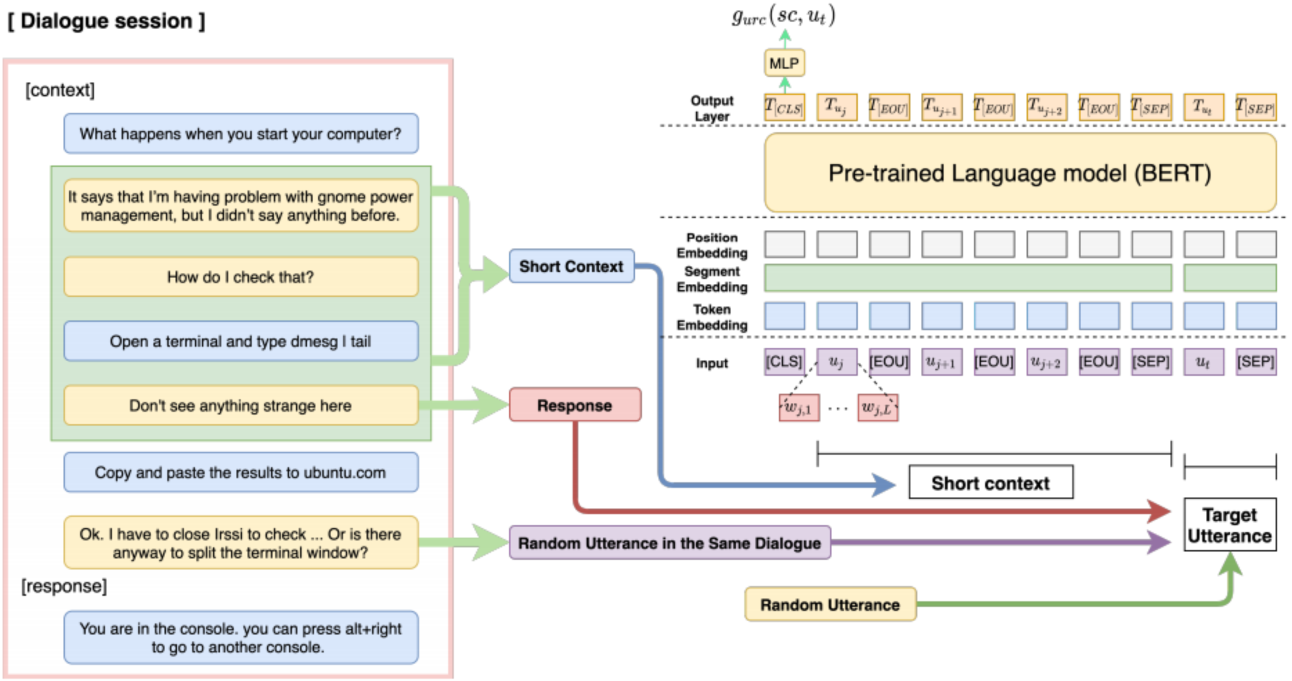


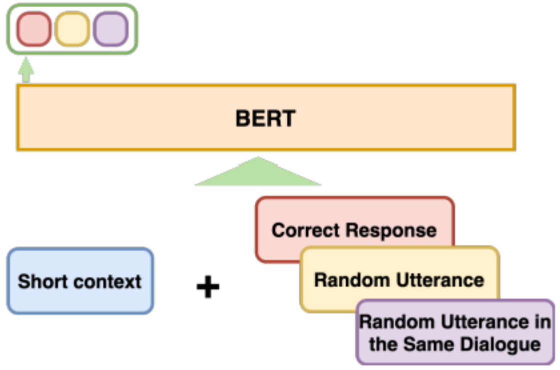
Fine-grained Post-training for Improving Retrieval-based Dialogue Systems

Janghoon Han[1,3], Taesuk Hong[1], Byoungjae Kim[1], Youngjoong Ko[2], Jungyun Seo[1]
[1]Department of Computer Science and Engineering, Sogang University
[2]Department of Computer Science and Engineering, Sungkyunkwan University
[3]LG AI Research
janghoon.han@lgresearch.ai
{hongtaesuk,wiz3021,seojy}@sogang.ac.kr
yjko@skku.edu

BERT-FP

- During the post-train, they convert dialog response selection task from **binary-classification (NSP)** to **three-classification**
  - Positive sample
  - Random negative sample
  - **Topic related hard negative sample** (utterance within the same session)

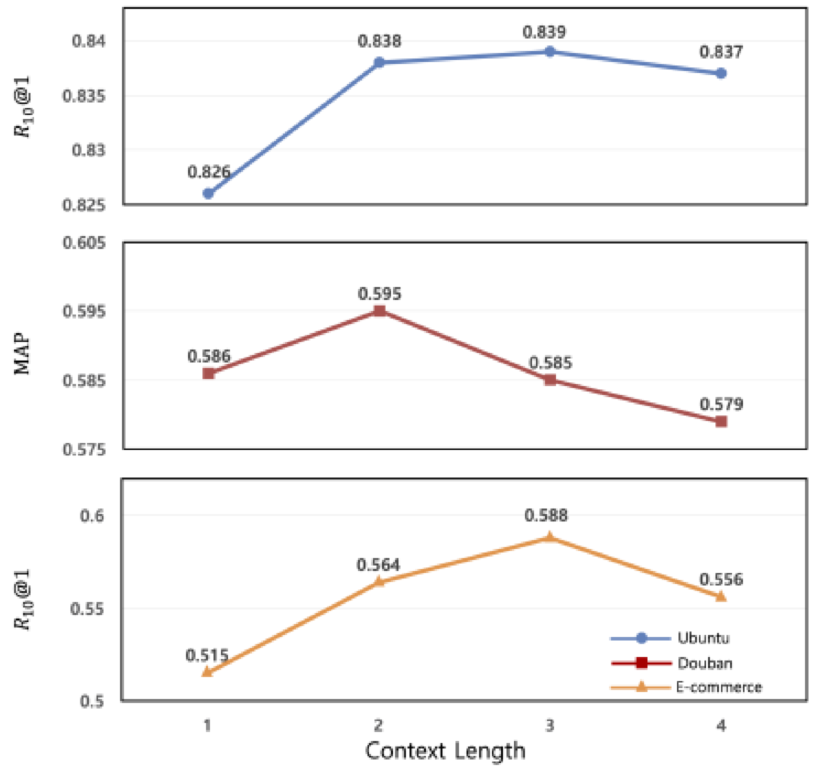




BERT-FP experiment

Models	Ubuntu			Douban						E-commerce		
	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	MAP	MRR	$P@1$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
TF-IDF (Lowe et al., 2015)	0.410	0.545	0.708	0.331	0.359	0.180	0.096	0.172	0.405	0.159	0.256	0.477
RNN (Lowe et al., 2015)	0.403	0.547	0.819	0.390	0.422	0.208	0.118	0.223	0.589	0.325	0.463	0.775
CNN (Kadlec et al., 2015)	0.549	0.684	0.896	0.417	0.440	0.226	0.121	0.252	0.647	0.328	0.515	0.792
LSTM (Kadlec et al., 2015)	0.638	0.784	0.949	0.485	0.537	0.320	0.187	0.343	0.720	0.365	0.536	0.828
SMN (Wu et al., 2017)	0.726	0.847	0.961	0.529	0.569	0.397	0.233	0.396	0.724	0.453	0.654	0.886
DUA (Zhang et al., 2018)	0.752	0.868	0.962	0.551	0.599	0.421	0.243	0.421	0.780	0.501	0.700	0.921
DAM(Zhou et al., 2018)	0.767	0.874	0.969	0.550	0.601	0.427	0.254	0.410	0.757	0.526	0.727	0.933
IOI (Tao et al., 2019)	0.796	0.894	0.974	0.573	0.621	0.444	0.269	0.451	0.786	0.563	0.768	0.950
ESIM (Chen and Wang, 2019)	0.796	0.894	0.975	-	-	-	-	-	-	0.570	0.767	0.948
MSN (Yuan et al., 2019)	0.800	0.899	0.978	0.587	0.632	0.470	0.295	0.452	0.788	0.606	0.770	0.937
BERT (Gu et al., 2020)	0.808	0.897	0.975	0.591	0.633	0.454	0.280	0.470	0.828	0.610	0.814	0.973
RoBERTa-SS-DA (Lu et al., 2020)	0.826	0.909	0.978	0.602	0.646	0.460	0.280	*0.495*	0.847	0.627	0.835	0.980
BERT-DPT (Whang et al., 2020)	0.851	0.924	0.984	-	-	-	-	-	-	-	-	-
BERT-VFT (Whang et al., 2020)	0.855	0.928	0.985	-	-	-	-	-	-	-	-	-
SA-BERT (Gu et al., 2020)	0.855	0.928	0.983	0.619	0.659	0.496	0.313	0.481	0.847	0.704	0.879	0.985
UMS_{BERT+} (Whang et al., 2021)	0.875	0.942	0.988	*0.625*	*0.664*	*0.499*	*0.318*	0.482	*0.858*	0.762	0.905	0.986
BERT-SL (Xu et al., 2021)	*0.884*	*0.946*	*0.990*	-	-	-	-	-	-	*0.776*	*0.919*	*0.991*
BERT-FP	**0.911**	**0.962**	**0.994**	**0.644**	**0.680**	**0.512**	**0.324**	**0.542**	**0.870**	**0.870**	**0.956**	**0.993**
(diff. %p)	(+2.7)	(+1.6)	(+0.4)	(+1.9)	(+1.6)	(+1.3)	(+0.6)	(+4.7)	(+1.2)	(+9.4)	(+3.7)	(+0.2)

BERT-FP achieve the SOTA performance, and significantly outperforms the previous works



The influence of the context length shows that the long conversation context may brings the noise for decision

BERT-FP conclusion

- Advantage
  - Their work demonstrates that post-train is still very important for this task
  - The analysis of the context length is interesting, which is rarely mentioned in previous works
- Disadvantage
  - There are still lots of training samples for post-train procedure, and the ablation study of this factor is missing.

Thanks !