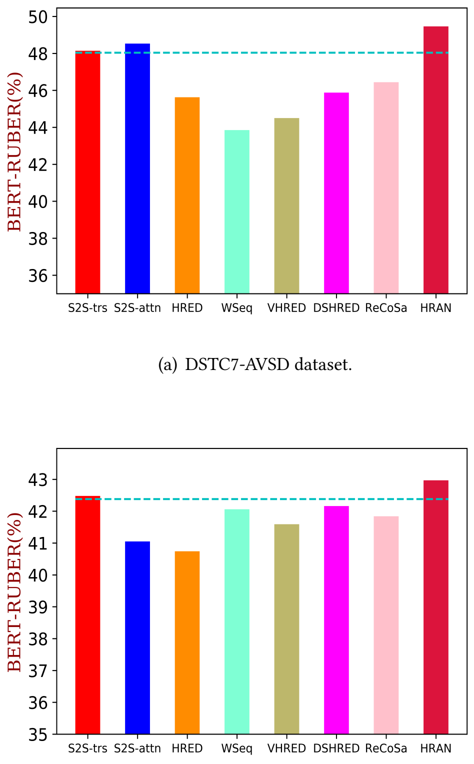
External Experiments

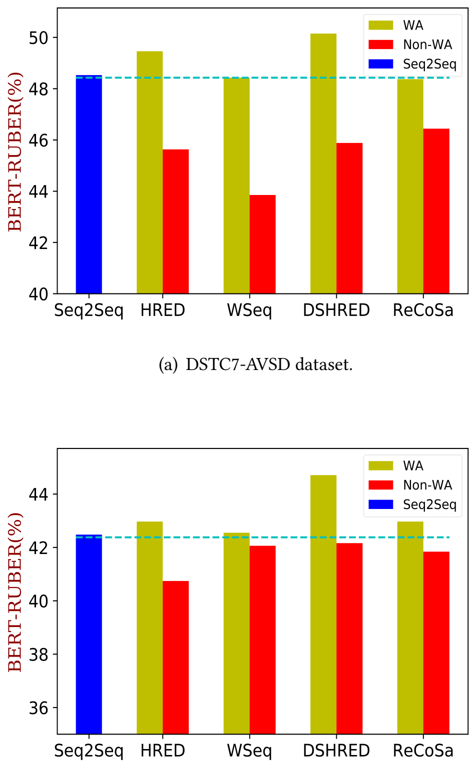


(a) DSTC7-AVSD dataset.

(b) PersonaChat dataset.

Figure 2: Hierarchical models vs. hierarchical models with word-level attention. The best Seq2Seq results are reported.

ACM Reference Format:
. 2018. External Experiments. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/1122445.1122456



(a) DSTC7-AVSD dataset.

(b) PersonaChat dataset.

Figure 1: Hierarchical models vs. hierarchical models with word-level attention on . The best Seq2Seq results are shown in blue bar.

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/10.1145/1122445.1122456

Model	BLEU-4	ROUGE-2	Dist-1	Dist-2	Average	Extrema	Greedy	BERTScore	RUBER
Seq2Seq+attn	1.93	3.04	0.72	4.57	69.67	83.97	52.30	13.57	44.63
Seq2Seq+trs	1.81	2.90	0.67	4.19	69.66	83.83	52.15	13.50	42.36
HRED+WA	1.83	2.86	0.95	6.02	69.22	83.79	52.04	13.05	48.04
WSeq+WA	1.84	3.04	0.73	4.36	69.06	83.70	52.03	13.64	45.65
DSHRED+WA	1.82	3.05	0.79	5.12	69.28	83.91	52.28	13.54	48.62
ReCoSa+WA	1.72	2.76	0.56	3.13	69.39	83.75	51.93	12.66	40.37

Table 1: Automatic evaluation on EmpChat dataset. Adding Word-level attention mechanism to hierarchical models. It should be noted that HRED+WA is the same as the HRAN model.

Model	BLEU-4	ROUGE-2	Dist-1	Dist-2	Average	Extrema	Greedy	BERTScore	RUBER
Seq2Seq+attn	2.66	4.67	0.79	4.57	63.36	83.52	48.87	16.41	41.05
Seq2Seq+trs	2.59	4.67	0.77	4.76	64.65	83.74	49.38	16.03	42.48
HRED	2.66	4.70	0.34	1.98	62.76	83.30	48.46	15.82	40.74
WSeq	2.54	4.56	0.41	2.41	63.25	83.53	48.43	15.96	42.06
VHRED	2.69	4.68	0.50	2.78	63.04	83.40	48.61	16.29	41.59
DSHRED	2.66	4.67	0.44	2.63	63.02	83.44	48.73	15.96	42.16
ReCoSa	2.72	4.70	0.47	2.93	63.29	83.31	48.56	15.65	41.84
HRAN	3.01	5.00	0.67	4.04	63.65	83.47	49.50	16.99	42.97

Table 2: Automatic evaluation (%) on PersonaChat dataset.

Model	BLEU-4	ROUGE-2	Dist-1	Dist-2	Average	Extrema	Greedy	BERTScore	RUBER
Seq2Seq+attn	2.66	4.67	0.79	4.57	63.36	83.52	48.87	16.41	41.05
Seq2Seq+trs	2.59	4.67	0.77	4.76	64.65	83.74	49.38	16.03	42.48
HRED+WA	3.01	5.00	0.67	4.04	63.65	82.47	49.50	16.99	42.97
WSeq+WA	2.66	4.60	0.59	3.20	62.94	83.43	48.69	16.52	42.55
DSHRED+WA	2.92	4.84	0.80	4.84	64.14	83.64	49.48	16.68	44.71
ReCoSa+WA	2.32	4.37	0.62	3.39	63.39	83.74	48.78	15.34	42.97

Table 3: Automatic evaluation on PersonaChat dataset. Adding Word-level attention mechanism to hierarchical models. It should be noted that HRED+WA is the same as the HRAN model.

Model	BLEU-4	ROUGE-2	Dist-1	Dist-2	Average	Extrema	Greedy	BERTScore	RUBER
Seq2Seq+attn	9.20	14.51	3.36	18.00	69.98	84.97	63.46	29.15	48.14
Seq2Seq+trs	9.33	14.32	3.22	17.08	69.71	84.85	63.82	28.75	48.53
HRED	8.92	13.42	2.92	14.81	68.54	84.14	61.79	27.69	45.63
WSeq	8.09	12.22	2.73	13.20	67.92	83.65	60.72	26.82	43.85
VHRED	8.13	12.41	2.82	14.00	67.74	83.78	60.58	26.65	44.50
DSHRED	7.96	12.69	2.93	15.37	69.01	84.31	61.54	27.73	45.88
ReCoSa	8.37	12.89	2.76	14.18	69.06	84.17	61.26	28.33	46.44
HRAN	9.22	14.36	3.58	19.43	69.75	84.93	63.73	29.33	49.46

Table 4: Automatic evaluation (%) on DSTC7-AVSD dataset.

Model	BLEU-4	ROUGE-2	Dist-1	Dist-2	Average	Extrema	Greedy	BERTScore	RUBER
Seq2Seq+attn	9.20	14.51	3.36	18.00	69.98	84.97	63.46	29.15	48.14
Seq2Seq+trs	9.33	14.32	3.22	17.08	69.71	84.85	63.82	28.75	48.53
HRED+WA	9.22	14.36	3.58	19.43	69.75	84.93	63.73	29.33	49.46
WSeq+WA	8.37	13.58	3.66	19.40	69.39	84.84	62.67	28.93	48.43
DSHRED+WA	9.20	14.11	3.88	20.51	69.34	84.94	63.44	29.13	50.15
ReCoSa+WA	9.25	14.09	3.41	17.24	69.44	84.63	63.25	29.21	48.37

Table 5: Automatic evaluation on DSTC7-AVSD dataset. Adding Word-level attention mechanism to hierarchical models. It should be noted that HRED+WA is the same as the HRAN model.