Semantic Noise Matters for Neural Natural Language Generation: Supplementary

TRAIN	TEST	System	Add	Miss	Wrong	InstOK
Original	Original	TGen−	2.8	686.2	4.8	192.2
		TGen	6.0	178.8	1.2	496.4
		TGen+	1.6	76.2	0.4	558.2
		SC-LSTM	121.6	823.6	426.2	7.8
Cleaned		TGen−	8.8	24.2	9.0	591.6
		TGen	4.2	0.8	0.2	624.8
		TGen+	1.0	0.2	0.2	628.6
		SC-LSTM	167.6	757.2	353.4	14.0
Cleaned missing		TGen−	6.0	98.2	9.4	525.2
		TGen	2.6	19.0	1.4	608.0
		TGen+	0.0	9.0	1.4	620.6
Cleaned added		TGen−	0.4	569.2	0.2	234.0
		TGen	2.0	132.2	0.2	501.6
		TGen+	0.2	62.8	0.2	567.0
Original	Cleaned	TGen−	39.4	1135.6	17.8	1089.4
		TGen	45.6	415.4	7.8	1469.8
		TGen+	23.6	230.2	5.2	1608.8
		SC-LSTM	858.6	1972.2	1057.6	39.0
Cleaned		TGen−	19.0	151.2	28.6	1667.8
		TGen	7.8	83.0	9.6	1751.4
		TGen+	1.8	72.6	7.0	1768.8
		SC-LSTM	876.2	1732.4	937.4	78.0
Cleaned missing		TGen−	49.4	328.4	30.0	1482.6
		TGen	42.8	162.0	10.8	1643.2
		TGen+	24.0	120.0	8.0	1702.8
Cleaned added		TGen−	2.2	1340.2	2.8	959.8
		TGen	6.0	373.6	1.8	1518.6
		TGen+	0.8	216.6	0.8	1646.2

Table 5: Absolute numbers of errors (added slots/missed slots/wrong slot values) and numbers of completely correct instances in all our experiments (compare to Tables 2 and 3 in the paper). Note that (1) the numbers are averages over 5 runs with different random network initializations, hence the non-integer values; (2) only numbers in the top half and the bottom half (with the same test set) are comparable. The original test set has 630 MRs and 4,352 slots in total. The cleaned test set has 1,847 MRs and 11,547 slots; however, for the runs with SC-LSTM these counts are 1,800 and 11,101, respectively, since some items had to be dropped due to preprocessing issues.