Supplementary Material for the Paper:
Are BLEU and Meaning Representation in Opposition?

Model	Size	Heads	MR	CR	SUBJ	MPQA	SST2	SST5	TREC	MRPC	SICK-E	SNLI	AvgAcc
Most frequent baseline			50.0	63.8	50.0	68.8	49.9	23.1	18.8	66.5	56.7	34.3	
Hill et al. en→fr†	2400	—	64.7	70.1	84.9	81.5	—	—	82.8	96.1	—	—	—
InferSent†	4096	—	81.1	86.3	92.4	90.2	84.6	—	88.2	76.2	86.3	84.5	—
InferSent	4096	—	81.5	86.7	92.7	90.6	85.0	45.8	88.2	76.6	86.4	83.7	81.7
GloVe-BOW	300	—	77.0	78.2	91.1	87.9	81.0	44.4	82.0	72.3	78.2	66.0	75.8
cs-FINAL-CTX	1000	—	68.7	77.4	88.5	85.5	73.0	38.2	88.6	71.8	82.1	70.2	74.4
cs-ATTN-ATTN	1000	1	68.2	76.0	86.9	84.9	72.0	35.7	89.0	70.7	80.8	69.3	73.4
cs-FINAL	1000	—	67.9	75.7	87.6	84.7	72.5	36.2	86.0	71.4	81.1	69.2	73.2
cs-MAXPOOL	1000	—	67.4	75.2	86.9	84.3	70.3	37.5	85.8	72.1	81.7	68.5	73.0
cs-AVGPOOL	1000	—	66.5	74.1	86.5	85.0	71.9	36.7	85.4	70.0	79.7	67.8	72.4
cs-ATTN-CTX	1000	4	66.5	74.8	85.7	84.7	70.1	36.1	88.2	70.4	79.5	66.0	72.2
cs-ATTN-ATTN	4000	4	64.9	72.7	84.3	85.1	70.1	33.5	88.8	69.7	78.0	65.2	71.2
cs-ATTN-ATTN	1000	4	64.0	72.6	84.6	84.2	67.9	33.2	89.0	70.1	78.0	64.6	70.8
cs-ATTN-ATTN	1000	8	62.9	71.7	83.6	84.2	67.0	34.2	86.2	69.8	76.6	63.2	70.0
de-MAXPOOL-CTX	600	—	60.0	69.2	77.0	73.1	61.4	32.4	80.2	70.7	78.8	68.0	67.1
de-ATTN-CTX	1200	12	61.1	70.0	77.3	71.7	63.5	32.4	78.4	69.8	77.4	65.0	66.7
de-ATTN-CTX	600	8	60.5	68.5	77.0	72.1	62.0	31.1	77.0	70.1	75.7	64.0	65.8
de-AVGPOOL-CTX	600	—	59.5	67.5	75.6	72.5	64.1	29.3	74.6	70.8	77.5	65.2	65.6
de-ATTN-CTX	600	12	59.7	68.4	77.0	71.2	61.2	30.9	78.0	71.1	76.0	61.9	65.5
de-FINAL	600	—	59.9	65.9	76.2	72.7	61.5	31.4	73.0	70.7	77.0	64.7	65.3
de-ATTN-CTX	600	3	60.3	67.0	75.4	72.7	60.6	30.4	77.0	69.9	76.0	63.3	65.3
de-ATTN-ATTN	600	1	60.0	66.5	72.8	72.2	61.7	29.5	74.2	70.5	76.9	63.8	64.8
de-ATTN-ATTN	600	3	60.7	67.5	74.1	71.8	60.6	30.1	75.0	69.5	74.7	61.5	64.5
de-FINAL-CTX	600	—	58.9	66.2	73.1	71.9	61.0	29.2	75.8	70.3	76.2	62.6	64.5
de-ATTN-ATTN	1200	6	58.7	65.9	75.4	72.3	61.0	29.7	78.4	70.1	72.3	59.6	64.3
de-TRF-ATTN-ATTN	600	3	58.8	64.9	76.2	71.7	60.3	30.4	72.0	71.2	72.5	61.4	63.9
de-ATTN-ATTN	1200	12	58.6	66.9	74.1	70.7	60.8	29.5	75.8	67.1	72.5	58.2	63.4
de-ATTN-ATTN	2400	12	57.4	66.0	74.0	70.9	58.5	27.7	76.0	67.7	73.9	59.8	63.2
de-TRF-ATTN-ATTN	2400	12	56.9	65.3	74.4	71.2	61.2	30.5	74.0	66.1	71.2	59.0	63.0
de-ATTN-ATTN	600	6	57.4	64.8	72.4	71.8	59.5	27.2	76.0	68.6	70.9	57.5	62.6
de-ATTN-ATTN	600	8	57.5	64.5	71.7	71.8	58.8	28.1	77.4	67.0	68.6	55.6	62.1
de-TRF-ATTN-ATTN	600	6	57.8	64.6	72.0	70.8	59.3	29.2	65.6	69.1	71.0	59.5	61.9
de-ATTN-ATTN	600	12	56.0	65.6	73.1	70.5	57.6	28.6	74.0	64.1	70.5	55.2	61.5
de-TRF-ATTN-ATTN	1200	12	56.6	64.9	71.4	71.0	56.7	29.6	66.2	67.9	68.8	58.2	61.1
de-ATTN-CTX	600	6	58.4	63.9	72.9	70.6	57.4	29.6	58.6	66.5	68.7	62.9	61.0
LM perplexity (*cs*)			1362.5	736.4	1059.0	3213.3	2099.1	1340.8	338.2	863.0	299.4	190.6	1150.2
% OOV (*cs*)			4.2	2.5	3.6	0.9	3.4	4.2	0.6	3.5	0.2	0.3	2.3
LM perplexity (*de*)			3776.8	2639.3	3137.7	8740.0	5003.3	3519.2	3790.8	5070.7	65.0	38.8	3578.2
% OOV (*de*)			22.8	13.1	21.0	27.9	24.4	23.3	16.7	25.6	1.7	1.5	17.8

Table S1: Classification accuracy of sentence representations on a number of SentEval tasks. Reprinted results are marked with †, others are our measurements. We give the out-of-vocabulary (OOV) rate and the perplexity of a 4-gram language model (LM) trained on the English side of the respective parallel corpus and evaluated on all available data for a given task. The last column is the average of each row.

Model	Size	Heads	SICK-R	STSB	STS12	STS13	STS14	STS15	STS16	AvgSim
InferSent	4096	—	.88/.83	.76/.75	.59/.60	.59/.59	.70/.67	.71/.72	.71/.73	.70
cs-MAXPOOL	1000	—	.81/.75	.72/.71	.52/.53	.47/.47	.54/.53	.61/.61	.58/.58	.60
cs-FINAL	1000	—	.80/.74	.74/.75	.54/.56	.42/.43	.55/.53	.60/.59	.55/.56	.60
cs-FINAL-CTX	1000	—	.82/.76	.74/.74	.51/.53	.44/.44	.52/.50	.62/.61	.57/.58	.60
GloVe-BOW	300	—	.80/.72	.64/.62	.52/.53	.50/.51	.55/.56	.56/.59	.51/.58	.59
cs-ATTN-ATTN	1000	1	.81/.76	.73/.73	.46/.49	.32/.33	.45/.44	.53/.52	.47/.48	.54
de-ATTN-CTX	1200	12	.76/.70	.52/.51	.46/.49	.31/.31	.50/.50	.58/.57	.51/.52	.52
de-ATTN-CTX	600	8	.75/.68	.52/.50	.47/.49	.30/.31	.52/.52	.56/.56	.48/.49	.51
de-ATTN-ATTN	600	1	.74/.67	.56/.55	.46/.48	.30/.31	.48/.48	.53/.53	.46/.47	.50
de-ATTN-CTX	600	3	.72/.65	.53/.52	.45/.48	.34/.35	.48/.48	.55/.54	.46/.46	.50
de-ATTN-CTX	600	12	.75/.68	.51/.49	.46/.47	.28/.29	.51/.50	.54/.54	.48/.48	.50
cs-AVGPOOL	1000	—	.78/.72	.70/.70	.47/.49	.29/.30	.38/.39	.44/.44	.41/.43	.50
de-MAXPOOL-CTX	600	—	.77/.71	.61/.60	.46/.48	.26/.28	.46/.46	.51/.52	.40/.42	.50
de-TRF-ATTN-ATTN	600	3	.70/.63	.53/.52	.47/.48	.31/.31	.47/.47	.52/.51	.47/.47	.49
de-AVGPOOL-CTX	600	—	.76/.69	.59/.58	.44/.46	.25/.27	.45/.45	.50/.49	.41/.42	.48
de-FINAL-CTX	600	—	.73/.66	.57/.55	.44/.47	.25/.27	.43/.43	.52/.51	.44/.44	.48
de-FINAL	600	—	.73/.66	.62/.60	.41/.44	.22/.24	.43/.43	.47/.47	.44/.44	.47
de-ATTN-ATTN	600	3	.67/.62	.50/.49	.44/.47	.27/.28	.43/.44	.50/.49	.45/.45	.47
de-TRF-ATTN-ATTN	2400	12	.66/.59	.50/.49	.42/.42	.28/.28	.46/.45	.51/.51	.44/.45	.46
de-TRF-ATTN-ATTN	1200	12	.61/.58	.51/.50	.44/.46	.26/.28	.43/.43	.50/.50	.47/.47	.46
de-TRF-ATTN-ATTN	600	6	.66/.59	.51/.49	.44/.45	.27/.28	.43/.43	.50/.51	.39/.41	.45
cs-ATTN-CTX	1000	4	.74/.70	.64/.64	.35/.38	.26/.27	.31/.31	.44/.44	.39/.40	.45
de-ATTN-ATTN	1200	12	.63/.58	.40/.39	.40/.43	.28/.29	.40/.41	.50/.49	.42/.41	.43
de-ATTN-CTX	600	6	.60/.57	.47/.47	.37/.38	.23/.26	.42/.43	.47/.48	.42/.44	.43
de-ATTN-ATTN	2400	12	.58/.59	.40/.39	.41/.44	.22/.25	.39/.39	.47/.47	.39/.38	.41
de-ATTN-ATTN	1200	6	.66/.60	.39/.39	.39/.42	.21/.23	.37/.37	.46/.45	.40/.39	.41
de-ATTN-ATTN	600	12	.59/.55	.40/.39	.39/.43	.24/.25	.37/.37	.46/.46	.39/.38	.40
de-ATTN-ATTN	600	6	.61/.56	.39/.38	.40/.43	.22/.23	.36/.36	.45/.45	.38/.37	.40
de-ATTN-ATTN	600	8	.57/.52	.37/.36	.38/.41	.24/.25	.35/.36	.46/.44	.38/.36	.39
cs-ATTN-ATTN	1000	4	.70/.66	.57/.56	.29/.32	.22/.21	.25/.25	.35/.35	.34/.34	.39
cs-ATTN-ATTN	4000	4	.72/.67	.57/.56	.29/.32	.22/.22	.24/.24	.36/.35	.32/.32	.39
cs-ATTN-ATTN	1000	8	.70/.65	.54/.52	.28/.31	.20/.20	.22/.22	.31/.32	.32/.33	.36
LM perplexity (*cs*)			299.4	1338.8	697.2	2783.9	1716.8	995.6	737.8	1224.2
% OOV (*cs*)			0.2	3.6	2.9	2.6	3.3	2.5	3.0	2.6
LM perplexity (*de*)			65.0	1301.4	1621.0	5041.8	2364.6	1096.7	2583.5	2010.6
% OOV (*de*)			1.7	19.6	18.5	23.5	19.9	13.3	17.2	16.2

Table S2: Similarity scores of sentence representations: Pearson/Spearman correlation between cosine similarity of pairs of sentence embeddings and similarity as marked manually by humans. "AvgSim" averages both correlation coefficients for all tasks. Perplexity and OOV rate as in Table S1.

Name	Size	Heads	Hy-Cl	Hy-NN (L_2/cos)	Hy-iDB	CO-Cl	CO-NN (L_2/cos)	CO-iDB	AvgPara
InferSent	4096	—	99.99	100.00/100.00	0.579	31.58	25.28/26.21	0.367	48.0
GloVe-BOW	300	—	99.94	100.00/100.00	0.654	34.28	20.29/19.72	0.352	46.9
cs-FINAL-CTX	1000	—	99.92	100.00/100.00	0.406	23.20	15.74/16.07	0.346	44.5
cs-MAXPOOL	1000	—	99.86	100.00/100.00	0.447	21.76	15.01/16.34	0.348	44.2
de-ATTN-CTX	600	8	98.11	99.86/ 99.90	0.348	21.64	15.40/17.32	0.349	44.1
cs-FINAL	1000	—	99.91	100.00/100.00	0.439	22.40	14.31/14.63	0.340	44.0
de-ATTN-CTX	1200	12	98.88	99.85/ 99.91	0.347	20.06	14.92/16.68	0.348	43.9
de-MAXPOOL-CTX	600	—	98.42	99.89/ 99.90	0.343	21.54	14.65/15.62	0.341	43.8
de-ATTN-CTX	600	3	97.81	99.77/ 99.87	0.328	19.74	15.28/16.43	0.343	43.7
de-ATTN-CTX	600	12	97.79	99.84/ 99.89	0.360	20.22	14.54/16.10	0.344	43.6
de-ATTN-CTX	600	6	98.11	99.79/ 99.86	0.358	20.44	14.48/15.57	0.342	43.6
de-ATTN-ATTN	600	1	97.70	99.71/ 99.73	0.352	19.74	14.95/16.26	0.340	43.6
de-AVGPOOL-CTX	600	—	97.72	99.59/ 99.60	0.312	20.04	13.49/14.27	0.337	43.2
cs-ATTN-ATTN	1000	1	99.88	99.91/ 99.91	0.347	21.54	11.15/11.50	0.331	43.1
de-ATTN-ATTN	600	3	97.42	99.64/ 99.75	0.314	17.36	13.35/14.35	0.333	42.8
de-FINAL	600	—	97.01	99.14/ 99.30	0.305	19.88	11.41/12.40	0.328	42.5
de-FINAL-CTX	600	—	96.65	99.66/ 99.70	0.323	17.22	12.06/12.84	0.333	42.3
de-TRF-ATTN-ATTN	600	3	95.79	99.61/ 99.64	0.315	15.76	13.20/14.04	0.340	42.3
cs-AVGPOOL	1000	—	99.80	99.99/ 99.99	0.387	17.90	8.36/ 8.61	0.311	41.9
de-ATTN-ATTN	1200	12	97.15	99.47/ 99.65	0.283	12.18	11.09/11.97	0.330	41.5
de-ATTN-ATTN	1200	6	98.05	99.74/ 99.80	0.289	11.90	9.84/10.69	0.327	41.3
de-ATTN-ATTN	2400	12	98.69	99.65/ 99.77	0.287	10.26	9.96/10.94	0.326	41.2
cs-ATTN-CTX	1000	4	99.75	99.72/ 99.74	0.287	14.60	7.52/ 7.54	0.318	41.2
de-ATTN-ATTN	600	6	96.03	99.62/ 99.71	0.287	12.22	9.92/10.59	0.323	41.1
de-TRF-ATTN-ATTN	2400	12	95.82	99.05/ 99.03	0.307	5.66	13.85/14.53	0.339	41.1
de-ATTN-ATTN	600	8	95.32	99.59/ 99.73	0.275	10.22	9.56/10.58	0.325	40.7
de-ATTN-ATTN	600	12	95.16	99.52/ 99.64	0.278	9.62	9.59/10.47	0.323	40.6
de-TRF-ATTN-ATTN	600	6	90.24	98.39/ 98.44	0.313	9.06	12.98/13.64	0.332	40.4
de-TRF-ATTN-ATTN	1200	12	90.71	98.21/ 98.22	0.301	7.06	13.10/13.70	0.333	40.2
cs-ATTN-ATTN	4000	4	99.54	98.89/ 98.98	0.252	11.52	5.54/ 5.51	0.303	40.1
cs-ATTN-ATTN	1000	4	99.26	98.90/ 98.93	0.253	10.84	5.16/ 5.20	0.299	39.9
cs-ATTN-ATTN	1000	8	99.41	98.17/ 98.09	0.243	10.24	4.51/ 4.64	0.287	39.4
LM perplexity / % OOV (*cs*)				668.5 / 1.2			238.5 / 0.1		
LM perplexity / % OOV (*de*)				3354.8 / 19.3			86.3 / 1.9		

Table S3: Evaluation by paraphrases on data from HyTER (Hy-) and COCO (CO-). "AvgPara" is simply the average of each row. Perplexity and OOV rate as in Table S1.