Supplementary Material

i **Training Configurations** We used an Nvidia Tesla K80 GPU to train the models. We divided the dataset into a training set and a validation set using 0.8:0.2 split. We evaluated the model while training and performed *early stopping* if the validation loss did not improve over ten evaluation steps. We used the parameter values mentioned in Table 3 and did not change it across the language pairs. For all the experiments we used the XLM-R large model.

Parameter	Value
learning rate	2e-5
maximum sequence length	128
number of epochs	3
adam epsilon	1e-8
warmup ratio	0.1
warmup steps	0
max grad norm	1.0
max seq. length	140
gradient accumulation steps	1

Table 3: Parameter Specifications.

ii **Hardware Specifications**

In Table 4 we mention the specifications of the GPU we used for the experiments of the paper.

Parameter	Value
GPU	Nvidia K80
GPU Memory	12GB
GPU Memory Clock	0.82GHz
Performance	4.1 TFLOPS
No. CPU Cores	2
RAM	12GB

Table 4: GPU Specifications.

iii **Other results**

In Table 5 and in Table 6 we show the F1-scores for gaps in target and for words in source. They follow the same format as Table 2. The Marmot baseline used in WMT 2018 does not support quality prediction for gaps in the target and words in the source. In addition, after WMT 2019, organisers did not release scores for gaps in target. For this reason, we do not report them in Table 5.

		IT			Pharmaceutical		
	Train Language(s)	En-Cs SMT	En-De NMT	En-De SMT	De-En SMT	En-LV NMT	En-Lv SMT
I	En-Cs SMT	0.2018	(-0.10)	(-0.08)	(-0.15)	(-0.02)	(-0.01)
	En-De NMT	(-0.17)	0.1672	(-0.07)	(-0.18)	(-0.01)	(-0.02)
	En-De SMT	(-0.08)	(-0.05)	0.4927	(-0.14)	(-0.06)	(-0.04)
	En-Ru NMT	(-0.14)	(-0.00)	(-0.15)	(-0.12)	(-0.01)	(-0.03)
	De-En SMT	(-0.18)	(-0.14)	(-0.33)	**0.4203**	(-0.29)	(-0.32)
	En-LV NMT	(-0.16)	(-0.09)	(-0.15)	(-0.12)	0.1664	(-0.01)
	En-Lv SMT	(-0.11)	(-0.12)	(-0.11)	(-0.16)	(-0.01)	0.2356
	En-De NMT	(-0.17)	(-0.01)	(-0.09)	(-0.14)	(-0.02)	(-0.04)
	En-Zh NMT	(-0.15)	(-0.08)	(-0.16)	(-0.16)	(-0.03)	(-0.06)
II	All	**0.2118**	**0.1773**	**0.5028**	0.4189	**0.1772**	**0.2388**
	All-1	(-0.03)	(-0.04)	(-0.08)	(-0.14)	(-0.01)	(-0.01)
III	Domain	0.2112	0.1695	0.4951	0.4132	0.1685	0.2370
IV	SMT/NMT	0.2110	0.1886	0.4921	0.4026	0.1671	0.2289
V	Marmot	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Best system	0.1671	0.1343	0.3161	0.3176	0.1598	0.1386

Table 5: GAP F1-Multi between the algorithm predictions and human annotations. Best results for each language by any method are marked in bold. Sections I, II and III indicate the different evaluation settings. Section IV shows the results of the state-of-the-art methods and the best system submitted for the language pair in that competition. **NR** implies that a particular result was *not reported* by the organisers. Zero-shot results are coloured in grey and the value shows the difference between the best result in that column for that language and itself.

		IT				Pharmaceutical			Wiki	
	Train Language(s)	En-Cs SMT	En-De NMT	En-De SMT	En-Ru NMT	De-En SMT	En-LV NMT	En-Lv SMT	En-De NMT	En-Zh NMT
I	En-Cs SMT	0.5327	(-0.08)	(-0.07)	(-0.09)	(-0.17)	(-0.02)	(-0.01)	(-0.12)	(-0.13)
	En-De NMT	(-0.17)	0.2957	(-0.07)	(-0.02)	(-0.19)	(-0.01)	(-0.02)	(-0.02)	(-0.08)
	En-De SMT	(-0.01)	(-0.05)	0.5269	(-0.67)	(-0.14)	(-0.06)	(-0.05)	(-0.08)	(-0.09)
	En-Ru NMT	(-0.14)	(-0.08)	(-0.18)	**0.5543**	(-0.14)	(-0.01)	(-0.03)	(-0.09)	(-0.08)
	De-En SMT	(-0.42)	(-0.21)	(-0.33)	(-0.31)	**0.4824**	(-0.29)	(-0.32)	(-0.23)	(-0.28)
	En-LV NMT	(-0.12)	(-0.09)	(-0.14)	(-0.03)	(-0.12)	0.4880	(-0.01)	(0.09)	(-0.08)
	En-Lv SMT	(-0.04)	(-0.16)	(-0.11)	(-0.09)	(-0.17)	(-0.02)	0.4945	(-0.15)	(-0.14)
	En-De NMT	(-0.11)	(-0.01)	(-0.08)	(-0.02)	(-0.15)	(-0.03)	(-0.04)	0.4456	(-0.06)
	En-Zh NMT	(-0.19)	(-0.08)	(-0.17)	(-0.03)	(-0.18)	(-0.05)	(-0.06)	(-0.07)	0.4040
II	All	**0.5442**	**0.3021**	**0.5445**	0.5535	0.4791	**0.4983**	**0.5005**	0.4483	0.4053
	All-1	(-0.02)	(-0.02)	(-0.06)	(-0.03)	(-0.16)	(-0.01)	(-0.01)	(-0.01)	(-0.04)
III	Domain	0.5421	0.2925	0.5421	0.5259	0.4672	0.4907	0.4991	0.4364	0.4021
IV	SMT/NMT	0.5412	0.2901	0.5412	0.5230	0.4670	0.4889	0.4932	0.4302	0.4012
V	Marmot	0.0000	0.0000	0.0000	NR	0.0000	0.0000	0.0000	NR	NR
	OpenKiwi	NR	NR	NR	0.2647	NR	NR	NR	0.3717	0.3729
	Best system	0.3937	0.2642	0.3368	0.4541	0.3200	0.3614	0.4945	**0.5672**	**0.4462**

Table 6: SOURCE F1-Multi between the algorithm predictions and human annotations. Best results for each language by any method are marked in bold. Rows I, II and III indicate the different evaluation settings. Row IV shows the results of the state-of-the-art methods and the best system submitted for the language pair in that competition. **NR** implies that a particular result was *not reported* by the organisers. Zero-shot results are coloured in grey and the value shows the difference between the best result in that column for that language and itself.