A Detailed WMT English-German Results

bitext	method	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**Avg**
160K	baseline	13.2	15.7	13.5	15.7	18.6	14.8	21.4	16.1
	SHARED	15.3	18.2	16.7	19.0	21.6	18.2	24.9	19.1
	SHARED+bitext-BPE	15.1	17.9	16.2	18.9	22.0	18.0	25.2	19.0
	SRC-ELMO	16.0	19.4	17.1	19.9	23.0	18.7	26.6	20.1
	SRC-FT	15.3	18.5	16.6	18.9	20.8	17.6	24.3	18.9
	TGT-ELMO	13.3	16.4	14.1	16.2	18.8	14.9	21.3	16.4
	TGT-FT	14.7	17.2	15.8	18.4	21.4	16.9	24.2	18.4
	SRC-ELMO+SHDEMB	17.4	20.8	18.6	21.5	24.9	20.3	29.0	**21.8**
320K	baseline	17.2	20.4	18.1	21.2	25.0	19.6	28.9	21.5
	SHARED	18.1	21.1	19.1	22.4	26.3	21.2	30.6	22.7
	SHARED+bitext-BPE	17.6	20.6	19.1	22.3	26.1	20.8	29.9	22.3
	src-elmo	18.8	22.3	21.1	24.0	27.5	22.2	32.5	24.1
	SRC-FT	19.0	22.5	20.9	23.5	26.9	22.2	32.1	23.9
	TGT-ELMO	16.7	20.7	18.2	20.9	24.1	19.4	28.0	21.1
	TGT-FT	16.1	19.4	17.1	20.0	23.1	18.5	26.3	20.1
	SRC-ELMO+SHDEMB	19.5	22.9	21.2	24.0	27.4	22.4	32.3	**24.2**
640K	baseline	19.2	22.9	21.2	24.5	27.9	22.4	33.1	24.5
	SHARED	19.9	23.4	22.1	25.1	28.8	23.0	34.1	25.2
	SHARED+bitext-BPE	19.4	22.8	21.7	24.9	28.4	22.9	33.6	24.8
	src-elmo	21.0	24.3	23.4	26.5	30.0	24.6	35.6	26.5
	SRC-FT	20.5	24.0	22.9	26.1	29.1	24.4	34.9	26.0
	TGT-ELMO	18.9	22.6	20.8	24.2	27.5	22.3	31.9	24.0
	TGT-FT	18.2	21.8	20.6	23.7	27.0	21.8	31.4	23.5
	SRC-ELMO+SHDEMB	21.2	25.1	23.9	26.7	30.2	24.7	36.2	**26.9**
1280K	baseline	20.9	24.6	23.6	26.5	30.5	24.7	36.2	26.7
	SHARED	21.1	24.6	24.6	27.6	31.0	25.2	37.3	27.3
	SHARED+bitext-BPE	20.5	24.0	23.9	26.2	30.6	24.8	36.2	26.6
	src-elmo	22.1	25.7	25.7	28.5	31.7	26.3	38.2	28.3
	SRC-FT	21.3	25.2	25.3	28.5	31.1	26.2	37.4	27.9
	TGT-ELMO	20.9	24.4	23.6	26.6	30.3	24.9	36.1	26.7
	TGT-FT	20.1	23.7	22.4	25.2	29.1	23.6	34.4	25.5
	SRC-ELMO+SHDEMB	22.3	26.0	26.3	28.9	32.6	26.8	38.6	**28.8**
2560K	baseline	21.7	25.6	25.4	28.2	32.3	26.2	39.1	28.4
	SHARED	22.2	25.9	25.7	28.3	32.1	26.3	38.9	28.5
	SHARED+bitext-BPE	21.8	25.5	25.5	27.9	32.1	26.0	38.6	28.2
	src-elmo	22.9	27.0	27.0	30.0	33.4	28.0	40.0	**29.8**
	SRC-FT	22.2	26.4	26.3	29.5	32.4	27.3	39.3	29.1
	TGT-ELMO	21.8	25.7	25.8	28.5	32.3	26.6	39.3	28.6
	TGT-FT	21.5	25.3	24.5	27.0	30.2	25.2	36.8	27.2
	SRC-ELMO+SHDEMB	23.1	27.2	27.1	29.7	33.7	27.9	40.0	**29.8**
5186K	baseline	23.1	26.8	27.7	30.1	33.6	27.9	40.1	29.9
	SHARED	22.6	26.6	27.7	30.5	33.4	28.2	40.2	29.9
	SHARED+bitext-BPE	22.5	26.0	27.0	29.7	33.4	27.7	40.6	29.6
	src-elmo	23.7	27.8	28.7	31.1	34.5	29.2	41.8	**31.0**
	SRC-FT	23.1	27.0	27.8	30.5	33.7	28.3	40.8	30.2
	TGT-ELMO	22.9	26.6	26.9	29.5	33.8	27.7	40.5	29.7
	TGT-FT	22.3	26.1	26.1	28.9	32.5	26.5	38.8	28.7
	SRC-ELMO+SHDEMB	23.4	28.0	28.8	31.2	34.5	28.7	41.8	30.9

Table 4: BLEU on newstest2012 to newstest2018 of WMT English-German translation in varioius simulated bitext size scenarios (cf. Figure 1).

B Training and inference speed

	train (tok/sec)	inference (tok/sec)
SHARED	528,802	2,334
SRC-ELMO	100,636	2,011
SRC-FT	57,753	2,080
TGT-ELMO	142,525	259
TGT-FT	95,313	299

Table 5: Training and inference speed of models trained on WMT English-German. Training speed based on 32 V100 GPUs. Inference speed measured on a single V100 and by batching up to 12K source or target tokens.