A Results with Standard Deviations

	news2016	news2017
DIR	39.0±0.1	34.3±0.1
DIR ENS	40.0±0.0	35.3±0.1
DIR+LM	39.8±0.1	35.2±0.3
CH+DIR+LM	41.0±0.0	36.2±0.2
- per word scores	40.0±0.0	35.1±0.2

Table 4: Online decoding accuracy for a direct model (DIR), ensembling two direct models (DIR ENS) and the channel approach (CH+DIR+LM). We ablate the impact of length normalization. Results are on news2017 of WMT De-En.

	5	10	50	100
DIR	39.1 ± 0.2	39.2 ± 0.0	39.3 ± 0.2	39.2 ± 0.1
DIR ENS	40.1 ± 0.2	40.2 ± 0.1	40.3 ± 0.2	40.3 ± 0.2
DIR+LM	40.0 ± 0.2	40.2 ± 0.1	40.6 ± 0.2	40.7 ± 0.1
DIR+RL	39.7 ± 0.1	40.1 ± 0.2	40.8 ± 0.2	40.8 ± 0.2
DIR+RL+LM	40.4 ± 0.2	40.9 ± 0.2	41.6 ± 0.2	41.8 ± 0.2
CH+DIR	39.7 ± 0.2	40.0 ± 0.2	40.5 ± 0.0	40.5 ± 0.1
CH+DIR+LM	40.8 ± 0.2	41.52 ± 0.1	42.8 ± 0.2	43.2 ± 0.0

Table 5: Re-ranking BLEU with different n-best list sizes on news2016 of WMT De-En.

	WMT De-En	WMT En-De	WMT Zh-En	IWSLT De-En
DIR	34.5 ± 0.2	28.4 ± 0.1	24.4 ± 0.1	33.3 ± 0.9
DIR ENS	35.5 ± 0.1	29.0 ± 0.1	25.2 ± 0.2	34.5 ± 0.3
DIR+LM	36.0 ± 0.2	29.4 ± 0.1	24.9 ± 0.3	34.2 ± 0.8
DIR+RL	35.7 ± 0.3	29.3 ± 0.0	25.3 ± 0.3	34.4 ± 0.6
DIR+RL+LM	36.8 ± 0.1	29.9 ± 0.1	25.4 ± 0.1	34.9 ± 0.6
CH+DIR	35.1 ± 0.1	28.3 ± 0.1	24.8 ± 0.2	34.0 ± 0.6
CH+DIR+LM	37.7 ± 0.1	30.5 ± 0.1	25.6 ± 0.1	35.5 ± 0.7

Table 6: Re-ranking accuracy with $k_1 = 50$ on four language directions on the respective test sets.