A Supplementary Material

A.1 Data and Preprocessing

We trained systems for four language pairs namely German-English, Arabic-English, Czech-English and Spanish-English using the data for the translation task of the International Workshop on Spoken Language Translation (Cettolo et al., 2014). Apart from using the in-domain TED corpus (≈ 200K sentences), we additionally used Europarl and News Corpus made available for the recent WMT campaign. For Arabic-to-English, we also news Corpus and a subset of UN corpus (1 Million sentences) (Eisele and Chen, 2010). We used a concatenation of dev- and test-2010 for tuning Neural MT models, test-2011 for development (tuning the Static Read and Write agent) and tests 2012-14 for testing. We used Moses (Koehn et al., 2007) preprocessing pipeline including tokenization and truecasing. For Arabic we used Farasa segmentation (Abdelali et al., 2016) with BPE (Sennrich et al., 2016) as suggested in (Sajjad et al., 2017a). We trained the BPE models separately for both the source and target datasets instead of jointly training limiting the number of operations to 49,500, as suggested in (Sennrich et al., 2016).

Pair	ID	Cat	test11	test12	test13	test14
ar-en	229K	1.26M	1199	1702	1169	1107
	4.4M	28.7M	22K	28K	24K	20K
	4.7M	30.2M	26K	32K	28K	24K
de-en	209K	2.4M	1433	1700	993	1305
	4.0M	61.9M	26K	29K	20K	24K
	4.3M	64.7M	27K	31K	21K	25K
cs-en	122K	900K	1013	1385	1327	–
	2.0M	20.3M	15K	21K	24K	–
	2.5M	23.6M	18K	25K	28K	–
es-en	188K	2.3M	1435	1385	–	–
	3.6M	66.9M	25K	27.5K	–	–
	3.8M	64.4M	27K	31K	–	–

Table 1: Data Statistics: First Row = Number of Sentences, Second Row: Number of Tokens in Source Language, Third Row: Number of Tokens in Target Language. First Column = statistics for the in-domain TED corpus, Second Column = Statistics for the Concatenated Data

A.2 Neural MT system

We train a 2-layer LSTM encoder-decoder with attention using the `seq2seq-attn` implementation (Kim, 2016) with the following settings: word vectors and LSTM states with 500 dimensions, SGD with an initial learning rate of 1.0, a decay rate of 0.5, and dropout rate of 0.3. The MT systems are trained for 13 epochs. We used uni-directional encoder because it is not possible to compute the encoder in right-to-left direction in the streaming scenario, due to unavailability of the full input sentence. Computing right-to-left encoder states with whatever input sequence is available is also not viable as it requires expensive recomputation after each input word is added.[6] We also trained the models by initializing the first decoder state with zeros, rather than using the final encoder state, which will not be available during stream decoding.

A.3 Average Proportion

In normal decoding, the BLEU metric is commonly used to calculate the quality of translations from a system. In stream decoding, we have to also consider the delay induced by the system along with its BLEU. In our work, we use *Average Proportion* (AP) as defined by Gu et al. (2017). AP is calculated as the total number of source words each target word required before being committed, normalized by the product of the source and target lengths. Formally, if $s(t_i)$ is the number of source words required for target word i before being committed, X is the source sequence and Y is the generated target sequence:

$$AP = \frac{1}{|X| \cdot |Y|} \sum_{t_i}^{Y} s(t_i) \qquad (1)$$

A.4 Incremental Decoder

Figure 4 shows the average results on the test-sets for the models trained on in-domain TED corpus. Here, we present the test-wise results for the interested reader. Missing table values correspond to unavailable test-sets on the IWSLT webpage. See Table 2.

A.5 Scalability

In section 4 we note that even though the `WIW` agent's performance is not significantly below our selected `STATIC-RW` agent, its AP is much higher. When we allow our `STATIC-RW` agent an AP similar to that of the `WIW` agent, we are able to restore the BLEU loss to be less than 1.5 for all

[6] Unlike left-to-right encoder which only requires single computation after each input word is added.

Pair	Agent	test12	test13	test14	Agent	test12	test13	test14
ar-en	WUE	30.16	28.16	25.53	WUE	32.84	32.23	28.95
	5, 2	29.31	27.72	25.21	7, 2	31.71	31.46	28.29
	WIW	28.06	25.86	23.75	WIW	29.48	28.82	26.52
	WID	19.89	17.24	15.64				
cs-en	WUE	22.95	25.03	–	WUE	27.97	30.50	–
	5, 4	22.97	24.46	–	8, 3	26.68	29.37	–
	WIW	21.78	21.99	–	WIW	25.20	27.43	–
	WID	16.37	17.07	–				
de-en	WUE	29.20	31.31	26.61	WUE	35.52	35.01	30.44
	6, 3	27.94	29.90	25.07	8, 3	28.62	31.71	27.09
	WIW	27.77	29.55	23.88	WIW	27.94	30.05	25.56
	WID	19.15	20.73	16.46				
es-en	WUE	29.65	–	–	WUE	32.78	–	–
	4, 1	29.04	–	–	8, 1	32.05	–	–
	WIW	28.65	–	–	WIW	30.59	–	–
	WID	21.90	–	–				

Table 2: Left Side: Test-wise results for ”Small” models in Figure 4, Right Side: Test-wise results for ”Large” models in Figure 4

.

language pairs except German-English. Here are the results in detail. See Table 2.