Appendix

A Model Training Parameters

Figure 1 shows the main training parameters for training multi-transformer architecture models with Marian. Other parameters not shown in the figure were kept as default. Figure 2 shows the parameters which were changed for training transformer-base models.

```
workspace: 8000
type: multi-transformer
dim-vocabs:   17181; 17181; 17181
dim-emb: 512
tied-embeddings-all: true
transformer-heads: 8
transformer-dim-ffn: 2048
transformer-ffn-depth: 2
transformer-ffn-activation: swish
transformer-dim-aan: 2048
transformer-aan-depth: 2
transformer-aan-activation: swish
transformer-decoder-autoreg: self-attention
transformer-preprocess: d
transformer-postprocess-emb: d
transformer-postprocess: dan
transformer-dropout: 0.1
transformer-dropout-attention: 0.1
transformer-dropout-ffn: 0.1
cost-type: ce-mean-words
max-length: 128
maxi-batch: 200
maxi-batch-sort: trg
optimizer: adam
optimizer-params:   0.9; 0.98; 1e-09
optimizer-delay: 8
sync-sgd: true
learn-rate: 0.0003
lr-decay: 0
lr-decay-start:   10; 1
lr-decay-freq: 5000
lr-decay-inv-sqrt: 16000
lr-warmup: 16000
lr-warmup-start-rate: 0
label-smoothing: 0.1
factor-weight: 1
clip-norm: 5
exponential-smoothing: 0.0001
beam-size: 6
normalize: 0.6
```

Figure 1: Main parameters for training multi-transformer models with Marian.

```
type: transformer
dim-vocabs:   17181; 17181
```

Figure 2: Different parameters for training baseline transformer models with Marian.

B Ablation Experiments with Dropout

Table 1 shows results from ablation experiments where we trained models for translation between English and Japanese and changed the values of *transformer-dropout*, *transformer-dropout-attention*, and *transformer-dropout-ffn* parameters between 0.1 (our main choice), 0.2 and 0.3 while keeping other parameters (including the seed - 347155) unchanged. This confirms that our initial choice was optimal.

	Dropout		
	0.10	**0.20**	**0.30**
Baseline	11.84	9.80	8.32
0-context	13.37	10.12	9.26
3-context-ood	14.43	9.44	6.17
3-context	14.90	9.43	6.48

Table 1: Ablation experiment results with different dropout values.