A Appendix

A.1 Data

NMT For DE→EN, we use *newstest2013* as the validation set, and use *newstest2017* as the test set. For DE→FR, we use *newstest2013* as the evaluation set, and use *newstest2012* as the test set.

A.2 Experimental Settings

NMT We implemented *RNNS2S* models with stacked bi-directional RNNs and implemented the self-attention in Transformer encoders to output the attention distributions. We use *Adam* (Kingma and Ba, 2015) as the optimizer. The initial learning rate is set to 0.0002. All the neural networks have 6 layers.[5] The size of embeddings and hidden units is 512. The attention mechanism in *Transformer* has 8 heads. We learn a joint BPE model with 32,000 subword units (Sennrich et al., 2016). All BLEU scores are computed with *SacreBLEU* (Post, 2018).

WSD Classification The classifiers are feed-forward neural networks with only one hidden layer, using ReLU non-linear activation. The size of the hidden layer is set to 512. We use Adam learning algorithm as well with mini-batches of size 3,000. The classifiers are trained using a cross-entropy loss. Each classifier is trained for 80 epochs[6] and the one performs best on the development set is selected for evaluation.

[5]The RNN encoder is a stack of three bi-directional RNNs which is equivalent to 6 uni-directional RNNs.

[6]The classifiers fed decoder states are trained 200 epochs to converge.