A Hyperparameters

For experiments on the News Commentary data we used 8000 BPE merges, whereas we used 16000 BPE merges for En–De experiments on the full dataset. For all the experiments, we used bidirectional GRUs and we set the embedding size to 256, we used word dropout with retain probability of 0.8 and edge dropout with the same probability, we used L2 regularization on all the parameters with value of 10^{-8}, translations are obtained using a greedy decoder. We placed residual connections (He et al., 2016) before every GCN layer. For the experiments on News Commentary data, we set GRU (for both encoder and decoder) and CNN hidden states to 512, we use Adam (Kingma and Ba, 2015) as optimizer with an initial learning rate of 0.0002, and we trained the models for 50 epochs. For large scale experiments on En–De, we set the GRU hidden states to 800, and instead of greedy decoding we employed beam search (beam 12). We trained the model for 20 epochs with the same hyperparameters.

B Datasets Statistics

	Train	Val.	Test
English–German	226822	2169	2999
English–German (full)	4500966	2169	2999

Table 5: The number of sentences in our datasets.

	Source	Target
English–German	37824	8099 (BPE)
English–German (full)	50000	16000 (BPE)

Table 6: Vocabulary sizes.