A Training Details

We use the original E2E dataset along with the default splits [1]. As for automatic evaluation, we use the tools provided in the E2E NLG Challenge [2].

We adopt Transformer base (Vaswani et al., 2017), with $d_{model} = 512$, $d_{hidden} = 2048$, $n_{heads} = 8$, $n_{layers} = 6$, $lr_{max} = 0.0005$, label-smooth=0.1, warmup=10000, dropout=0.3, weight-decay=0.01. Source and target side share embeddings. All the models are trained using 1 Nvidia Tesla V100 GPU with the batch size of 8000 tokens, maximum 70K steps for training or finetune on task-related data, and 1.5M steps for pre-training on open-domain. All the hyperparameters follow the default setting of the LevT paper (Gu et al., 2019) except for the training steps. We manually examine the performance of 10K, 30K, 50K, 70K, 90K, 110K training steps of the text stitch model trained on the free text of E2E training data, and find 70K performs best in METEOR. Then, we use 70K for all settings. As for pre-training, we choose the training steps by examined performances 1M, 1.5M, 2M steps. The whole model has about 60M parameters. We use the fairseq (Ott et al., 2019) framework, and it takes about 7 seconds to finish 100 steps.

B Annotation Rules

We randomly select 300 data from the E2E test set for human evaluation, and collect the generated texts of each system on these 300 data. Two paid annotators are asked to annotate the generated texts of each system. They discuss with each other to resolve conflict annotations.

We follow the *fluency* definition of Ferreira et al. (2019) that the sentence is fluent when it is grammatical and flow in a natural, easy to read manner. As for *dropped slots* and *modified slots*, we follow the definition of Puzikov and Gurevych (2018). They refer to the situation that the generated text dropped certain slots or modified the value of certain slots in the input data, respectively.

Since there can exist up to 8 slot-value pairs in the input data, it is very time-consuming to manually examine the dropped-slot errors and modified-slot errors for all data. Therefore, we first examine if the input slot values are present in the output texts. Then, only those sentences with missing slot values are sent to annotators for further examination. Pilot experiments show that this process introduces no extra errors.

C Detailed Templates

See supplementary materials.

[1] https://github.com/tuetschek/e2e-dataset
[2] https://github.com/tuetschek/e2e-metrics