## Appendix

The generative model is an LSTM with two layers with hidden dimension equals 500, input dimension equals 400, and dropout of 0.2. The inference network uses a one-layer Bi-LSTM with hidden size of 500 and input size of 400 to encode the sentence. We use large max segment length, $L = 8$ (segmental for data-to-text) and $L = 1$ (linear chain for POS induction) and 0.2 dropout in the inference network. The Bi-LSTM used for encoding the source table is has hidden dimension of 300. Both the generative model and the inference network share word embeddings.

The batch size is 10 for WikiBio and 20 for PTB and E2E. The generative model and the inference network are optimized by Adam (Kingma and Ba, 2014) gradient clipping at 1, with learning rate of $0.002$ and $0.001$ respectively. Parameters are all initialized from a standard Gaussian distribution. The learning rate decays by a factor of two for any epoch without improvement of loss function on validation set, and this decay condition is not triggered until the eighth epoch for sufficient training. Training is done for max of 30 epochs and allows for early stopping.

For data-to-text problem, we need to encode the data table. We encode the E2E source table by directly concatenating word embeddings and field embeddings and indices for each token, for example, if the word $w$ is the $i$th token from left and $j$th token from right under field type $f$, then we represent the token using a concatenation $[\text{emb}(w) \cdot \text{emb}(f) \cdot \text{emb}(i) \cdot \text{emb}(j)]$. We encode the WikiBio table by passing a bidirectional-LSTM through the tokens in the table, where each token has similar embedding by concatenation as above. The encoding of the table is denoted as $c$. We use copy attention (Gu et al., 2016; Gulcehre et al., 2016) in the generative model, and the attention vector $\alpha$ at a time step is parametrized by the class label $z$ at that time step. Recall the contextual representation is $\sum_i \alpha_i \cdot c_i$, where $\alpha_i = \text{softmax}(\text{score}(h_t, c_i))$ and $\text{score}(h_t, c_i) = (W_z(h_t) + b_z) \cdot (W_2(c_i) + b_2)$, the parametrization from $z$ happens during the feedforward network indexed by $z$. For the WikiBio data, we use a dual attention mechanism described in (Liu et al., 2018), where the first attention is the same as above and the second attention uses a different encoder context $c'_i$, the $c'_i$ only looks at the concatenation of field type and field index, but not the field value itself, i.e. $[\text{emb}(f) \cdot \text{emb}(i) \cdot \text{emb}(j)]$. Then the two attention forms two different sets of $\alpha_i$ and they are multiplied together and renormalized to form an attention.