Appendix A

Training details

Our models are trained with early stopping by running the proposed evaluation method on the development set after every epoch. We use a single layer LSTM for the encoder and decoder. We tie the embeddings of the encoder and the decoder, and preliminary experiments showed similar results without tying. In all models the size of the embedding vectors is similar to the size of the LSTM units (128/256/512). We decode using beam search with a beam size of 12. All model parameters, including the embeddings are randomly initialized and learned during training. For optimization we use SGD with an initial learning rate of 1.0 and decay the learning rate by 0.5 when there is no improvement on the validation set when completing an epoch.