Generating long and coherent text is an important but challenging task, particularly for open-ended language generation tasks such as story generation. Despite the success in modeling intra-sentence coherence, existing generation models (e.g., BART) still struggle to maintain a coherent event sequence throughout the generated text. We conjecture that this is because of the difficulty for the decoder to capture the high-level semantics and discourse structures in the context beyond token-level co-occurrence. In this paper, we propose a long text generation model, which can represent the prefix sentences at sentence level and discourse level in the decoding process. To this end, we propose two pretraining objectives to learn the representations by predicting inter-sentence semantic similarity and distinguishing between normal and shuffled sentence orders. Extensive experiments show that our model can generate more coherent texts than state-of-the-art baselines.