Learning Neural Templates for Text Generation

Encoder-Decoder architectures: The Encoder module give a representation of the souce text. The decoder module interprets this representation to generate text. Existing architectures are 1. Uninterpretable (blackbox) and 2. Difficult to control the ouput generation (style, content, etc). It is difficult to relate how a word in the output is related to a word in the input.

Related Work: Earlier work in Natural Language generation, have made use of hand-engineered templates. Work has been done to automate these template generation, using clustering methods on similar sentences and then abstracting templated fields.

Proposed Approach: A novel architecture, that uses a Hidden Semi Markov Model decoder (HSMM) that jointly learns the latent templates and output text. According to the authors, it provides a principled approach to template-like text generation. The authors maximize the log-marginal and use RNNs to parameterize a discrete emission distribution

Method: According to standard definitoin, a *record* in a knowledge base (KG) comprises of *type*, *entity*, and *value*. A *template* is a sequence of typed text segments, with parts of the segments filled in by the knowledge base. An HSMM specifies a joint distribution on the observations and latent segments. There are two discrete variables specified for this distribution. l_t is a length variable, that specifies the length of the current segment, and f_t a binary variable that indicates wether segment finishes at time step t. The factorized distribution is given by

$$p(y, z, l, f|x; \theta) = \prod_{t=0}^{T-1} p(z_{t+1}, l_{t+1}|z_t, l_t, x)^{f_t} \times \prod_{t=1}^{T} p(y_{t-l_t+1:t}|z_t, l_t, x)^{f_t} \tag{1}$$

The model is fatorized into, **Transition** distribution, **Length** distribution, and **Emission** distribution.

$p(z_{t+1}|z_t)$ represents the transition distribution. To formulate this, the authors use matrices $\mathbf{A}, \mathbf{B}$ as state embeddings and $\mathbf{C}, \mathbf{D}$ as parameterized non-linear functions. The length distrubution is uniform. The emission distribution is modelled as an RNN decoder.

Learning: The variables z, l and f are unobserved. They are margnialized over, and maximum log-marginal likelihood is used, by applying standard backpropagation.

Conclusion: This method allows for diversity in generation and provides interpretable states.