Appendix

A Structured prediction formulation

He et al. proposed to incorporate such structural dependencies at decoding time by augmenting the loglikelihood function with penalization terms for constraint violations

$$\mathbf{f}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \log p(y_i|\mathbf{x}) - \sum_{c \in C} c(\mathbf{x}, \mathbf{y}_{1:i}) \quad (7)$$

where, each constraint function c applies a non-negative penalty given the input $\mathbf{x}$ and a length-t prefix $\mathbf{y}_{1:t}$.