Local attention

1. **Find the most relevant position $\mathbf{a_j}$ in the source**

- Monotonic alignments: $a_j = j$
- Predictive alignments: $a_j = I \cdot \sigma(b^T \tanh(W s_j))$

2. **Attend only positions within a window $[\mathrm{a_j} - \mathrm{h}; \mathrm{a_j} + \mathrm{h}]$**

- Compute scores as usual
- Probably multiply by a Gaussian centered in a_j

Luong et. al. Effective Approaches to Attention-based Neural Machine Translation, 2015.