* Runtime scales linearly w/ size of decoder vocabulary

Seq2seq Models (Encoder - Decoder)

- Can view many tasks as mapping from an input sequence of tokens to an output series of tokens

- Syntactic Parsing

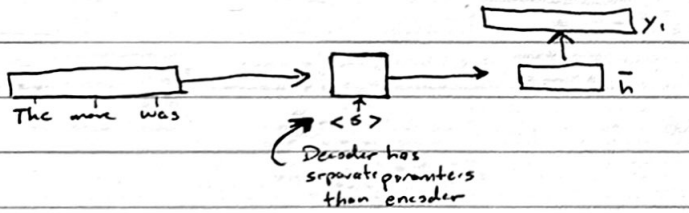
  The Dog Ran → (S (NP (DT The) (NN Dog)) (VP (VBD Ran)))

- Semantic Parsing

  What state borders texas → $\lambda x \; state(x) \wedge borders(x, e89)$

* Slightly different than language modeling (decoder-only) b/c the input and output vocabularies can be different.

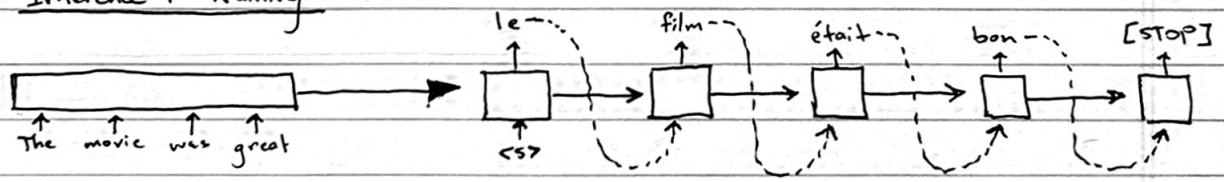
- Seq2seq models : generate next word conditioned on previous output as well as input
  - W is |vocab| × |hidden state|, softmax over entire vocabulary



$$P(y_i \mid x, y_1 \cdots y_{i-1}) = \text{softmax}(W\bar{h})$$

$$P(y \mid x) = \prod_{i=1}^{n} P(y_i \mid x, y_1 \cdots y_{i-1})$$

- Inference + Training



  - Inference : Need to compute argmax over the word predictions and then feed that to the next transformer cell
  - Decoder is advanced one state at a time until [STOP] is reached
  - Encoder can just be run a single time

* Training : Same as LM training, maximize probability of the gold sequence y (now conditioned on input x)