Can't we just use the chain rule?

- Can't we just use the chain rule of probability to factor the joint distribution into a distribution over its suffix conditioned on its prefix?

$$P(sentence) = P(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}) = \underbrace{P(\mathbf{x}^{(1)})}_{\text{yes}} \cdot \underbrace{P(\mathbf{x}^{(2)} \,|\, \mathbf{x}^{(1)})}_{\text{yes}} \cdot \underbrace{P(\mathbf{x}^{(3)} \,|\, \mathbf{x}^{(1)}, \mathbf{x}^{(2)})}_{\text{yes}} \ldots \underbrace{P(\mathbf{x}^{(T)} \,|\, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T-1)})}_{\text{No!}}$$

- No because we end up with (essentially) the same problem, now it's too many possible sequences over $T - 1$ tokens.

- We need to make a simplifying assumption regarding the independence of words in a sentence.

The Markov approximation

- From previous slide, notice that can estimate ***some*** of the factors of the joint distribution:

$$P(sentence) = P(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}) = \underbrace{P(\mathbf{x}^{(1)})}_{\text{yes}} \cdot \underbrace{P(\mathbf{x}^{(2)} \mid \mathbf{x}^{(1)})}_{\text{yes}} \cdot \underbrace{P(\mathbf{x}^{(3)} \mid \mathbf{x}^{(1)}, \mathbf{x}^{(2)})}_{\text{yes}} \ldots \underbrace{P(\mathbf{x}^{(T)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T-1)})}_{\text{No!}}$$

- The Markov assumption: $P(\mathbf{x}^{(t)} \mid \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t-1)}) = P(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-n)}, \ldots, \mathbf{x}^{(t-1)})$

- Example:

$$P(dog \mid the, quick, brown, fox, jumped, over, the, lazy) = P(dog) \quad \leftarrow \quad (n = 1)$$

$$P(dog \mid the, quick, brown, fox, jumped, over, the, lazy) = P(dog \mid lazy) \quad \leftarrow \quad (n = 2)$$

$$P(dog \mid the, quick, brown, fox, jumped, over, the, lazy) = P(dog \mid the, lazy) \quad \leftarrow \quad (n = 3)$$

$$P(dog \mid the, quick, brown, fox, jumped, over, the, lazy) = P(dog \mid over, the, lazy) \quad \leftarrow \quad (n = 4)$$