n-gram models

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

- n-gram models are language models that use the Markov assumption with a specific selection of n.

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

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

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

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

- This should look familiar to Lecture-03 (Word2Vec)! Skip-gram modeling generalizes *n*-grams by not considering the sequence order of the *context* words.

Estimating *n*-gram probabilities

- The maximum likelihood estimate of an *n*-gram model:

$$P(\mathbf{x}^{(t)} \,|\, \mathbf{x}^{(t-n)}, \ldots, \mathbf{x}^{(t-1)}) \overset{MLE}{=} \frac{count\left(\mathbf{x}^{(t-n)}, \ldots, \mathbf{x}^{(t)}\right)}{\sum_{j=1}^{N} count\left(\mathbf{x}^{(t-n)}, \ldots, \mathbf{x}_j^{(t)}\right)}$$

Example from Jurafsky & Martin

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>
```

Here are the calculations for some of the bigram probabilities from this corpus

$P(\texttt{I}|\texttt{<s>}) = \frac{2}{3} = .67$ $\quad P(\texttt{Sam}|\texttt{<s>}) = \frac{1}{3} = .33$ $\quad P(\texttt{am}|\texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>}|\texttt{Sam}) = \frac{1}{2} = 0.5$ $\quad P(\texttt{Sam}|\texttt{am}) = \frac{1}{2} = .5$ $\quad P(\texttt{do}|\texttt{I}) = \frac{1}{3} = .33$