NLP

N-grams And Regular Languages

- N-grams are just one way to represent weighted regular languages
- More about this in the lecture on regular languages

Generative Models

- Unigram: generate a word, then generate the next one, until you generate </S>.

- Bigram: generate <S>, generate a word, then generate the next one based on the previous one, etc., until you generate </S>.

Engineering Trick

- The MLE values are often on the order of 10^{-6} or less
  - Multiplying 20 such values gives a number on the order of 10^{-120}
  - This leads to underflow
- Use (base 10) logarithms instead
  - 10^{-6} becomes -6
  - Use sums instead of products

NLP