More Sophisticated LMs

- **N-gram language models**
  - In general, $p(w_1 w_2 \dots w_n) = p(w_1)p(w_2|w_1)\dots p(w_n|w_1 \dots w_{n-1})$
  - n-gram: conditioned only on the past n-1 words
  - E.g., bigram: $p(w_1 \dots w_n) = p(w_1)p(w_2|w_1) \, p(w_3|w_2) \dots p(w_n|w_{n-1})$
- **Exponential language models** (e.g., Maximum Entropy model)
  - P(w|history) as a function with features defined on "(w, history)"
  - Features are weighted with parameters (fewer parameters!)
- **Structured language models**: generate text based a latent (linguistic) structure (e.g., probabilistic context-free grammar)
- **Neural language models** (e.g., recurrent neural networks, word embedding): model p(w|history) as a neural network