- Introduction to generative models of language
  - What are they?
  - Why they're important
  - Issues for counting words
  - Statistics of natural language
  - **Unsmoothed n-gram models**

Administrative stuff

- Course website:
  - http://www.cs.cornell.edu/courses/cs4740/2015fa/
- Piazza
  - Signup link:
    - piazza.com/cornell/fall2015/cs4740
  - Class link:
    - piazza.com/cornell/fall2015/cs4740/home
- CMS: turn ON notifications

Critiques

- How to do well on them
  - do not summarize the paper
  - do identify one or two points about the paper to discuss in more depth, e.g. a paragraph for each
- Due on CMS on Weds@11:59pm
- ALSO due in hardcopy form AT THE BEGINNING OF THE NEXT CLASS
- Grading: check (A-/B+), check+ (A), check- (C)
- Late assignments: 1/2 grade late per weekday

Collaboration on Critiques

- We **encourage** discussion of the paper with others in the class — in person and via Piazza, etc.
- The content of the critique must be yours and the writing of the critique must be done by you

***From last class

- How many word types are in this text?

Marseille is a dog. He might do a dog trick later.

[Assumptions: punctuation is treated as a word; capitalized and lowercase versions of words are treated the same.]

A. 10
B. 11
C. 12
D. 13
E. something else

***From last class

- How many word tokens are in this text?

Marseille is a dog. He might do a dog trick later.

[Assumptions: punctuation is treated as a word; capitalized and lowercase versions of words are treated the same.]

A. 10
B. 11
C. 12
D. 13
E. something else

Goals

- Determine the next word in a sequence
  - Probability distribution across all words in the language
  - $P\ (w_n \mid w_1\ w_2 \ldots w_{n-1})$

- Determine the probability of a sequence of words
  - $P\ (w_1\ w_2 \ldots w_{n-1}\ w_n)$

***From last class

- We are studying n-gram models of word prediction. A *bigram* model is bases decisions on:
  A. the preceding word
  B. the two preceding words
  C. the preceding and following word
  D. no words in the context at all
  E. I have no idea.

(possible) Models of word prediction

- Simplest model
  - Let any word follow any other word (equally likely)
    - » E.g., P $(w_n \mid w_{n-1})$ —> P $(w_n \mid w_{n-1})$ —>

      P (word *n follows* word *n-1*)

- Probability distribution at least obeys actual relative word frequencies
    - » P $(w_n \mid w_{n-1})$ —>

Probability of a word sequence

- $P\ (w_1\ w_2\ \ldots\ w_{n-1}\ w_n)$

- Problem?
- Solution: *approximate* the probability of a word given all the previous words…

N-gram approximations

- Markov assumption: probability of some future event (next word) depends only on a limited history of preceding events (previous words)
- Bigram model

$$P(w_n|w_1^{n-1}) \approx P(w_n|w_{n-1})$$ predict next word

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-1})$$ prob of a word sequence

- Trigram model
  – Conditions on the two preceding words
- N-gram approximation

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1})$$ prob of a word sequence

Probability of a word sequence: bigram estimation

$P(w_1\ w_2 \ldots w_{n-1}\ w_n)$

$$P(w_1^n) = P(w_1)\ P(w_2|w_1)\ P(w_3|w_1^2)\ \ldots P(w_n|w_1^{n-1})$$

$$= P(w_1)\ \ P(w_2|w_1)\ \ P(w_3|w_2)\ \ \ldots\ \ P(w_n|w_{n-1})$$

$$= P(w_1|\langle s\rangle)\ \ P(w_2|w_1)\ \ P(w_3|w_2)\ \ \ldots\ \ P(w_n|w_{n-1})$$

$$= \prod_{k=1}^{n} P(w_k|w_1^{k-1})$$