Lecture 9: Statistical inference

19 Jan 2004

n-gram models

“Nothing happens in a vacuum”
“Past behavior to guess about future...”

Statistical Inference
Past → Future Words are seen — to return about
Taking in large corpora → to reason about other text, language

p (W = w — history) What word is next in the text given the conditioning on “history”
If we go with a fine-grained history, it can be more discriminatory.
Speaker + contact + medium + ... = good guess
Too fine-grained will be unreliable (not good stats)
Coarse-grained is reliable

n-gram models * not the same as an “engram”
predict word n based on word1 to word n-1
p ( Wn — W1 ..., Wn-1)

Simplification to make: sentences are independent
not always valid (“Dow Jones” in one sentence—what is the chance it will be in the next?)
Finding first word of the sentence What about long sentences? Running into sparse data Position dependent (i.e. x, y, z; 3, 4, 5)

Markov Assumption
Only the previous n events matter

0th-level Markov assumption/unigram
Where each word randomly generated

1st order Markov model / bigram
Each only depends on the previous
Consists of pairs of words
Tri-grams, 4-grams...
Effects of having larger n-gram models

- Limited amount of data, RAM
- Large models too sparse?
More accuracy
Subject-verb agreement can be expressed
Long distance dependencies sought

Trigrams doing alright performance-wise

wears his coat
wears his coat
eats his lunch

\[ p \text{ (coat — his)} = \frac{2}{3} \]
\[ p \text{ (lunch — his)} = \frac{1}{3} \]
\[ p \text{ (Wn = lunch — Wn-1 = his)} \]
\[ p \text{ (his — wears)} = 1 \]
\[ p \text{ (his — eats)} = 1 \]

\{eats his coat, wears his lunch\}
These two cases are not bloody likely in the English language!

Trigram models will capture those longer dependencies