Dealing with unseen data

Arne Köhn

December 10, 2013

What happens with unseen data?

- We assume that everything has been observed
- X unseen $\Rightarrow \forall s \in States.P(X|s) = 0$
- No meaningful computation possible if X observed

An example

assume the following HMM:

- $\chi = \{A, B\}$
- π: $\{A \Rightarrow 1, B \Rightarrow 0\}$
- Only observation: “foo”
- What happens if we observe “bar”?

Estimation of unseen Events

- Rare events can be unseen by chance
- Seeing an event once could also be due to chance
- Simple approach: add one to every possible observation
- Result: unseen events have half the probability of events seen once
- More elaborate: Good-Turing

Transitions can also be unseen!

- Especially when using higher-order Markov Models
- suppose $P(C|BA) = 0$ according to our model
- We don't want that!
- Smoothing to the rescue: Use e.g.
  $0.95 * P(C|BA) + 0.04 * P(C|B) + 0.01 * P(C)$
- Factors need to be set to something that works