Using Rejuvenation to improve Particle Filtering for Bayesian Word Segmentation

Benjamin Börschinger[*,+] Mark Johnson[*]

[*]Macquarie University, [+]Heidelberg University

July 10, 2012

Outline

Bayesian Word Segmentation

Particle Filtering

Rejuvenation

Evaluation

Word Segmentation

- breaking speech into smaller units (e.g. words)

$$\text{j} \,_{\vartriangle}\, \text{u} \,_{\blacktriangle}\, \text{w} \,_{\vartriangle}\, \text{ɑ} \,_{\vartriangle}\, \text{n} \,_{\vartriangle}\, \text{t} \,_{\blacktriangle}\, \text{t} \,_{\vartriangle}\, \text{u} \,_{\blacktriangle}\, \text{s} \,_{\vartriangle}\, \text{i} \,_{\blacktriangle}\, \text{ð} \,_{\vartriangle}\, \text{ə} \,_{\blacktriangle}\, \text{b} \,_{\vartriangle}\, \text{ʊ} \,_{\vartriangle}\, \text{k}$$

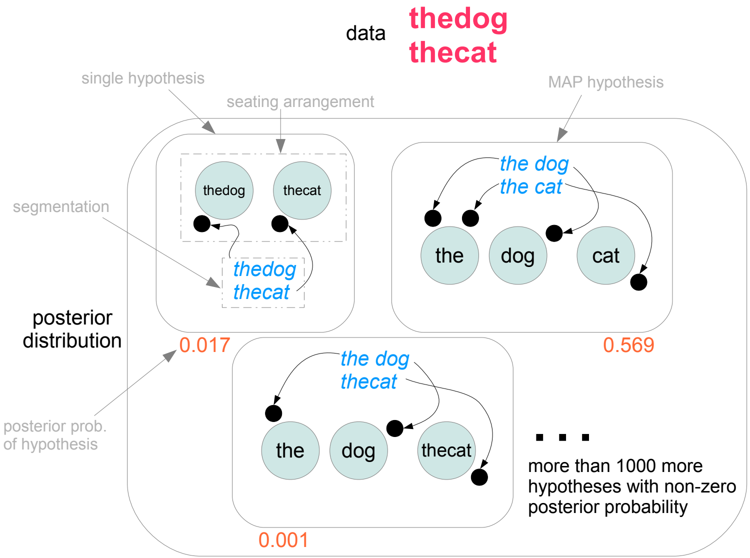
"you want to see the book"

- "learning to put boundaries at the right places"
- Goldwater introduced **non-parametric Bayesian segmentation models** building on the Dirichlet Process
- assign a probability to every sequence of words ⇒ define a posterior distribution overs segmentations for any given sequence of segments

The Goldwater Model for Word Segmentation

- infinite number of possible words, but only expect to observe a few
- $\Rightarrow$ model underlying lexicon G as draw from a Dirichlet Process
  - a distribution over all possible words
  - but mass concentrated on a (relatively) small subset
- integrating out the lexicon gives rise to a **Chinese Restaurant Process**
- just need to store a **seating arrangement** for previous word tokens instead of explicitly representing the "infinite" G

Inference



data
thedog
thecat
single hypothesis
seating arrangement
MAP hypothesis
segmentation
thedog
thecat
thedog
thecat
the dog
the cat
the
dog
cat
posterior
distribution
0.017
0.569
posterior prob.
of hypothesis
the dog
thecat
the
dog
thecat
. . .
more than 1000 more hypotheses with non-zero posterior probability
0.001

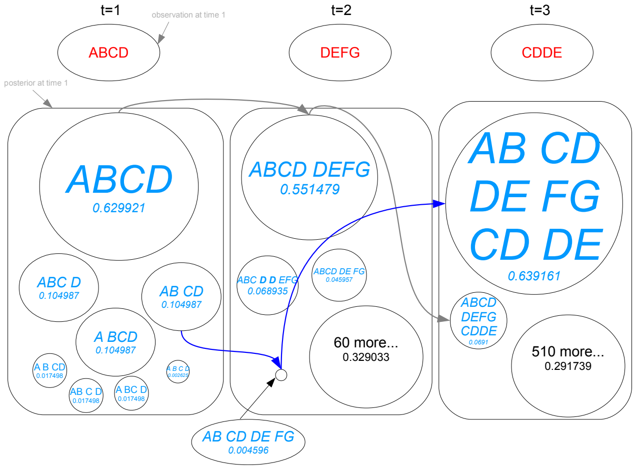
Particle Filtering for Word Segmentation

- infeasible to determine posterior exactly $\Rightarrow$ approximations
- SISR Particle Filter is asymptotically correct **online** inference algorithm
  - "make use of observations one at a time, [...] and then discard them before the next observations are used" (Bishop 2006: 73)
- maintains multiple weighted hypotheses (= particles) and updates these incrementally
- each particles corresponds to specific seating arrangement that summarizes previous segmentation choices
- described in Börschinger and Johnson, 2011

Problems for Particle Filtering

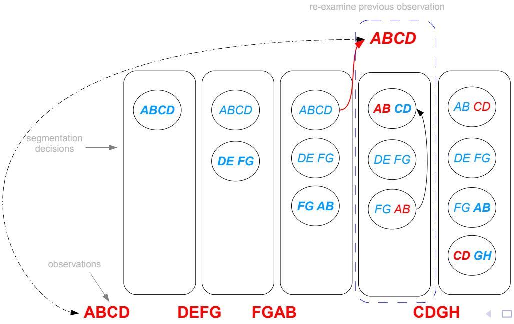
- "make use of observations one at a time, [...] and **then discard them** before the next observations are used" (Bishop 2006:73)
- $\Rightarrow$ once you made a decision, you can't really change it
  - exponential number of possibilities
  - "errors" propagate
  - later evidence may be relevant for evaluation of early evidence [example next slide]

Problems for Particle Filtering, Illustration



Addressing the problem - Rejuvenation

- using more and more particles? $\Rightarrow$ practical limitations (and loss of cognitive plausibility)
- relax the online constraint $\Rightarrow$ **Rejuvenation** (Canini et al. 2009)
  - given current knowledge, see if "better" alternatives to previous analyses now available
  - $\Rightarrow$ re-analyse fixed number of randomly chosen previous observations



Rejuvenation

- after each utterance, for each particle
  - do N times
    - randomly choose previously observed utterance
    - remove words "learned" from that utterance from particle
    - sample novel segmentation for utterance, given modified state and add new analysis back in
- can use sampling method also used in utterance based MCMC sampler (Mochihashi et al., 2009)
  - $\Rightarrow$ doesn't affect asymptotic guarantee
  - if we do (too) many rejuvenation samples, at last utterance turns into batch sampler
- requires storage of previous observations $\Rightarrow$ not strictly online
- but still **incremental** $\Rightarrow$ processes evidence as it becomes available

Evaluation

- evaluate on de-facto standard, Bernstein-Ratner corpus as per (Brent 1999)
  - 9790 phonemically transcribed utterances of child directed speech
- focus on Bigram model (Unigram model in paper)
- compare 1- and 16-particle filter with 100 rejuvenation steps to
  - "original" (online) particle filters (Börschinger and Johnson, 2011), including a 1000-particle filter
  - utterance-based ("ideal") batch sampler (with annealing)
  - 1-particle filter with 1600 rejuvenation steps (vs 16-particle filter w. 100)

Evaluation

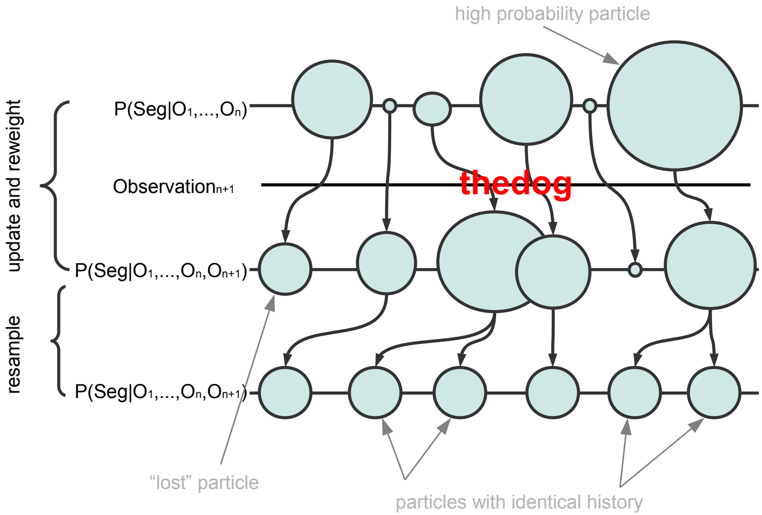
- online particle filters have low Token F-scores
- 1-particle filter with rejuvenation outperforms all online particle filters
- with 16 particles, performance similar to batch sampler
- 1-particle filter with 1600 rejuvenation steps outperforms batch sampler

Learner	TF
MHS	70.93 ($\sim$ Goldwater results)
Online-PF_{1}	49.43
Online-PF_{16}	50.14
Online-PF_{1000}	57.88
Rejuv-$PF_{1,100}$	66.88
Rejuv-$PF_{16,100}$	70.05
Rejuv-$PF_{1,1600}$	**74.47**

Conclusion and outlook

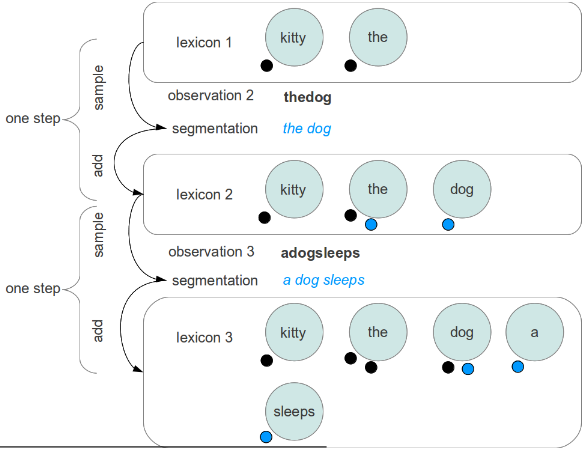
- Rejuvenation considerably boosts particle filter performance...
- ...but requires storage of observations
- in the future:
    - exploring variants of rejuvenation, i.e.
        - only remembering a fixed number of observations
        - choosing previous observations according to their recency (Pearl et al. 2011)
        - only rejuvenating at certain intervals
        - adapting the number of rejuvenation steps
        - ...
    - making the models more realistic (phonotactics, ...)
    - applying particle filters to other tasks (Adaptor Grammars)

Particle Filtering for Word Segmentation



Updating an individual Particle

- each particle is a lexicon (cum grano salis[1])
- updating a lexicon corresponds to
  - sampling a segmentation given the current lexicon
  - adding the words in this segmentation to the lexicon



[1] more precisely: a seating arrangement

Evaluation, inference

- what about **inference** performance?
- compare log-probability of training data at end
- particle filters with rejuvenation much better than without but still considerable gap
- even the Bigram model seems to benefit from "biased" search (see also Pearl et al. (2011))
- suspect that batch samplers suffer from too much data due to spurious "global" generalizations

Learner	TF	log-probability ($\times 10^3$)
MHS	70.93	**-237.24**
Online-PF_1	49.43	-265.40
Online-PF_{16}	50.14	-262.34
Online-PF_{1000}	57.88	-254.17
Rejuv-$PF_{1,100}$	66.88	-257.65
Rejuv-$PF_{16,100}$	70.05	-251.66
Rejuv-$PF_{1,1600}$	**74.47**	-249.78