From sounds to words:

Bayesian modelling of early language acquisition
[Excerpted and annotated by Henry S. Thompson 15 March 2013]

Sharon Goldwater

ilcc | Institute for Language, Cognition and Computation

THE UNIVERSITY of EDINBURGH informatics

Ohio State University, 25 May 2012

Bayesian learning

- The Bayesian learner seeks to identify an explanatory linguistic hypothesis that
  - accounts for the observed data.
  - conforms to prior expectations.

$$\underbrace{P(h|d)}_{\text{posterior}} \propto \underbrace{P(d|h)}_{\text{likelihood}}\underbrace{P(h)}_{\text{prior}}$$

- Focus is on the goal of computation, not the procedure (algorithm) used to achieve the goal.
  - As in Marr's layering of computation-algorithm-implementation

Data:

lookatthedoggie
seethedoggie
shelookssofriendly
…

Hypotheses:

lookatthedoggie
seethedoggie
shelookssofriendly
…

l o o k a t t h e d o g g i e
s e e t h e d o g g i e
s h e l o o k s s o f r i e n d l y
…

look at thed oggi e
se e thed oggi e
sh e look ssofri e ndly
…

look at the doggie
see the doggie
she looks so friendly
…

$P(d|h)=1$

i like pizza
what about you
…

abc def gh
ijklmn opqrst uvwx
…

$P(d|h)=0$

Bayesian segmentation

- In the domain of segmentation, we have:
  - Data: unsegmented corpus (transcriptions).
  - Hypotheses: sequences of word tokens.

$$\underbrace{P(h|d)}_{\text{posterior}} \propto \underbrace{P(d|h)}_{\text{likelihood}} \underbrace{P(h)}_{\text{prior}}$$

= 1 if concatenating words forms corpus,
= 0 otherwise.

Encodes assumptions of learner.

- Optimal solution is the segmentation with highest prior probability.
  - Because the likelihood is just a binary switch

Bayesian model

Assumes word w_i is generated as follows:

1. Is w_i a novel lexical item?

$$P(yes) = \frac{\alpha}{n+\alpha}$$

Fewer word types = Higher probability

$$P(no) = \frac{n}{n+\alpha}$$

[n is the number of words (types) we've learned]

[α is a model parameter, in practice around 100]

[Note that the above correctly mean that at the *very* beginning, when n is 0, p(yes) == 1 and p(no)==0]

Bayesian model

Assume word w_i is generated as follows:

2. If novel, generate phonemic form $x_1 \ldots x_m$:

$$P(w_i = x_1 \ldots x_m) = \prod_{i=1}^{m} P(x_i)$$

Shorter words = Higher probability

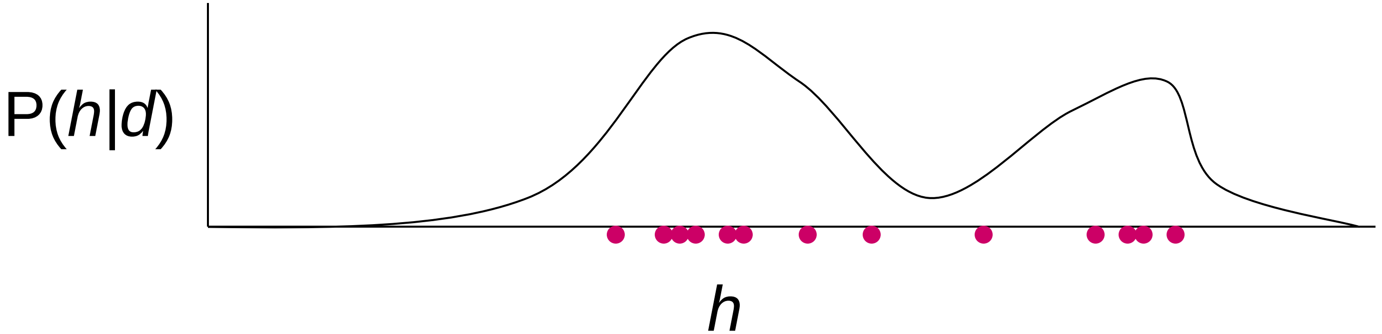
If not, choose lexical identity of w_i from previously occurring words:

$$P(w_i = w) = \frac{n_w}{n}$$

Power law = Higher probability [the rich get richer and the poor stay poor]

Learning algorithm

- Model defines a distribution over hypotheses. We use Gibbs sampling to find a good hypothesis.
  - Iterative procedure produces samples from the posterior distribution of hypotheses.



  - A batch algorithm, assumes perfect memory for data.
- A kind of **Monte Carlo** algorithm
  - Intelligent semi-random hill-climbing

Unigram model: simulations

- Same corpus as Brent (Bernstein-Ratner, 1987):
  - 9790 utterances of phonemically transcribed child-directed speech (19-23 months).
  - Average utterance length: 3.4 words.
  - Average word length: 2.9 phonemes.
- Example input:

```
yuwanttusiD6bUk
lUkD*z6b7wIThIzh&t
&nd6dOgi
yuwanttulUk&tDIs
...
```

Results

- Example segmentation:

```
youwant to see thebook
look theres aboy with his hat
and adoggie
you wantto lookatthis
lookatthis
havea drink
okay now
whatsthis
whatsthat
whatisit
look canyou take itout
...
```

What happened?

- Model assumes (falsely) that words have the same probability regardless of context.

$P(\texttt{that}) = .024 \qquad P(\texttt{that}|\texttt{whats}) = .46 \qquad P(\texttt{that}|\texttt{to}) = .0019$

- Positing amalgams allows the model to capture word-to-word dependencies.
- Empirical and theoretical analysis: undersegmentation is the optimal solution for any (reasonable) unigram model.

Results after extension to bigram prior

- Example segmentation:

```
you want to see the book
look theres a boy with his hat
and a doggie
you want to lookat this
lookat this
have a drink
okay now
whats this
whats that
whatis it
look canyou take it out
...
```

Summary

- More sophisticated use of available statistical information leads to better segmentation.
- Good segmentations of naturalistic data can be found using fairly weak prior assumptions.
  - Utterances are composed of discrete units (words).
  - Units tend to be short.
  - Some units occur frequently, most do not.
  - Unit boundaries have properties distinct (at least to some extent) from unit internals.