Online Learning Mechanisms for Bayesian Models of Word Segmentation

Sharon Goldwater
School of Informatics
University of Edinburgh

Lisa Pearl   Mark Steyvers
Department of Cognitive Sciences
University of California, Irvine

Workshop on Psychocomputational Models of Human Language Acquisition
VU University Amsterdam, 29 July 2009

Language acquisition as induction

Input

(specific linguistic observations)

Abstract internal representation/generalization

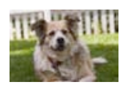
Output

(specific linguistic productions)

Bayesian modeling: ideal vs. constrained

- Typically an ideal observer approach asks what the optimal solution to the induction problem is, given particular assumptions about representation and available information.
- Here we investigate constrained learners that implement ideal learners in cognitively plausible ways.
  - How might limitations on memory and processing affect learning?

Word segmentation



- Given a corpus of fluent speech or text (no utterance-internal word boundaries), we want to identify the words.

whatsthat thedoggie yeah wheresthedoggie	→	whats that the doggie yeah wheres the doggie

Word segmentation

- One of the first problems infants must solve when learning language.
- Infants make use of many different cues.
  - Phonotactics, allophonic variation, metrical (stress) patterns, effects of coarticulation, and statistical regularities in syllable sequences.
- Statistics may provide initial bootstrapping.
  - Used very early (Thiessen & Saffran, 2003)
  - Language-independent, so doesn't require children to know some words already

Bayesian learning

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

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

- Ideal learner: Focus is on the goal of computation, not the procedure (algorithm) used to achieve the goal.
- Constrained learner: Uses same probabilistic model, but algorithm reflects how humans might implement the computation.

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 or biases in the learner.

- Optimal solution is the segmentation with highest prior probability.

An ideal Bayesian learner for word segmentation

- Model considers hypothesis space of segmentations, preferring those where
  - The lexicon is relatively small.
  - Words are relatively short.
- The learner has a perfect memory for the data
  - Order of data presentation doesn't matter.
  - The entire corpus (or equivalent) is available in memory.
- Note: only counts of lexicon items are required to compute highest probability segmentation. (ask us how!)

Goldwater, Griffiths, and Johnson (2007, 2009)

Investigating learner assumptions

- If a learner assumes that words are independent units, what is learned from realistic data? [unigram model]
- What if the learner assumes that words are units that help predict other units? [bigram model]

Approach of Goldwater, Griffiths, & Johnson (2007): use a Bayesian ideal observer to examine the consequences of making these different assumptions.

Corpus: child-directed speech samples

- Bernstein-Ratner corpus:
  - 9790 utterances of phonemically transcribed child-directed speech (19-23 months), 33399 tokens and 1321 unique types.
  - Average utterance length: 3.4 words
  - Average word length: 2.9 phonemes
- Example input:

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

≈

```
youwanttoseethebook
looktheresaboywithhishat
andadoggie
youwanttolookatthis
...
```

Results: Ideal learner

Precision: #correct / #found

Recall: #found / #true

	Word Tokens Prec	Word Tokens Rec	Boundaries Prec	Boundaries Rec	Lexicon Prec	Lexicon Rec
Ideal (unigram)	61.7	47.1	**92.7**	61.6	55.1	**66.0**
Ideal (bigram)	**74.6**	**68.4**	90.4	**79.8**	**63.3**	62.6

- The assumption that words predict other words is good: bigram model generally has superior performance
- Both models tend to undersegment, though the bigram model does so less (boundary precision > boundary recall)

Results: Ideal learner sample segmentations

Unigram model

```
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
...
```

Bigram model

```
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
...
```

How about online learners?

- Online learners use the same probabilistic model, but process the data incrementally (one utterance at a time), rather than in a batch.
  - Dynamic Programming with Maximization (DPM)
  - Dynamic Programming with Sampling (DPS)
  - Decayed Markov Chain Monte Carlo (DMCMC)

Considering human limitations

What is the most direct translation of the ideal learner to an online learner that must process utterances one at a time?

Dynamic Programming: Maximization

For each utterance:
- Use dynamic programming to compute probabilites of all segmentations, given the current lexicon.
- Choose the best segmentation.
- Add counts of segmented words to lexicon.

	you want to see the book
0.33	yu want tusi D6bUk
0.21	yu wanttusi D6bUk
0.15	yuwant tusi D6 bUk
…	…

- Algorithm used by Brent (1999), with different model.

Considering human limitations

What if humans don't always choose the most probable hypothesis, but instead sample among the different hypotheses available?

Dynamic Programming: Sampling

For each utterance:
- Use dynamic programming to compute probabilites of all segmentations, given the current lexicon.
- Sample a segmentation.
- Add counts of segmented words to lexicon.

	you want to see the book
0.33	yu want tusi D6bUk
0.21	yu wanttusi D6bUk
0.15	yuwant tusi D6 bUk
…	…

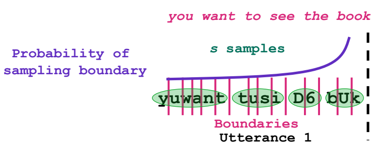
- Particle filter: more particles ⟺ more memory

Considering human limitations

What if humans are more likely to sample potential word boundaries that they have heard more recently (decaying memory = recency effect)?

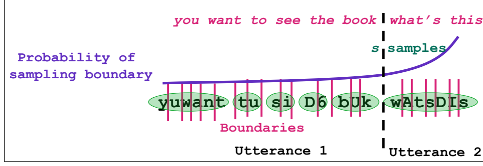
Decayed Markov Chain Monte Carlo

For each utterance:
- Probabilistically sample *s* boundaries from all utterances encountered so far.
- Prob(sample *b*) = b_a^{-d} where b_a is the number of potential boundary locations between *b* and the end of the current utterance and *d* is the decay rate (Marthi et al. 2002).
- Update lexicon after the *s* samples are completed.



Decayed Markov Chain Monte Carlo

For each utterance:
- Probabilistically sample *s* boundaries from all utterances encountered so far.
- Prob(sample *b*) = b_a^{-d} where b_a is the number of potential boundary locations between *b* and the end of the current utterance and *d* is the decay rate (Marthi et al. 2002).
- Update lexicon after the *s* samples are completed.

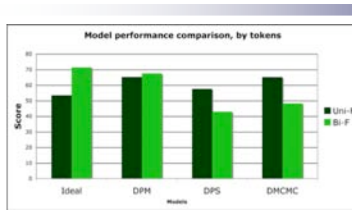


Decayed Markov Chain Monte Carlo

Decay rates tested: 2, 1.5, 1, 0.75, 0.5, 0.25

	Probability of sampling within current utterance
d = 2	.942
d = 1.5	.772
d = 1	.323
d = 0.75	.125
d = 0.5	.036
d = 0.25	.009

Results: unigrams vs. bigrams



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

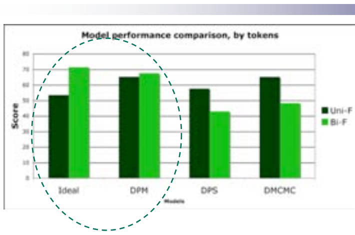
Precision:
#correct / #found

Recall:
#found / #true

Results from 2nd half of corpus

DMCMC Unigram: d=1, s=10000
DMCMC Bigram: d=0.5, s=15000

Results: unigrams vs. bigrams



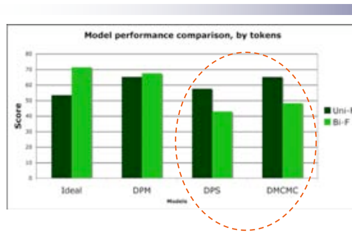
$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:
#correct / #found

Recall:
#found / #true

Like the Ideal learner, the DPM bigram learner performs better than the unigram learner, though improvement is not as great as in the Ideal learner. The bigram assumption is helpful.

Results: unigrams vs. bigrams



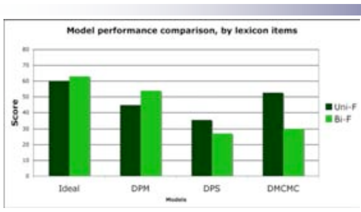
$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:
#correct / #found

Recall:
#found / #true

However, the DPS and DMCMC bigram learners perform worse than the unigram learners. The bigram assumption is not helpful.

Results: unigrams vs. bigrams for the lexicon



Results from 2nd half of corpus

$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

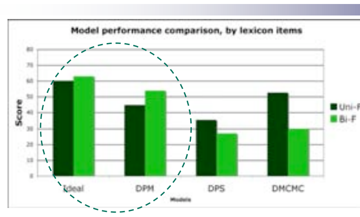
#correct / #found

Recall:

#found / #true

Lexicon = a seed pool of words for children to use to figure out language-dependent word segmentation strategies.

Results: unigrams vs. bigrams for the lexicon



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

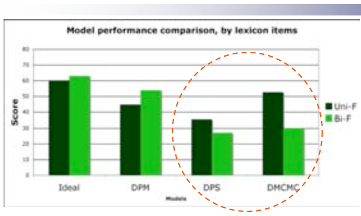
#correct / #found

Recall:

#found / #true

Like the Ideal learner, the DPM bigram learner yields a more reliable lexicon than the unigram learner.

Results: unigrams vs. bigrams for the lexicon



$$F = \frac{2 * Prec * Rec}{Prec + Rec}$$

Precision:

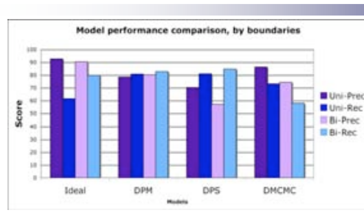
#correct / #found

Recall:

#found / #true

However, the DPS and DMCMC bigram learners yield much less reliable lexicons than the unigram learners.

Results: under vs. oversegmentation



Results from 2nd half of corpus

Precision:

#correct / #found

Recall:

#found / #true

Undersegmentation: boundary precision > boundary recall

Oversegmentation: boundary precision < boundary recall

Results: under vs. oversegmentation

Precision:

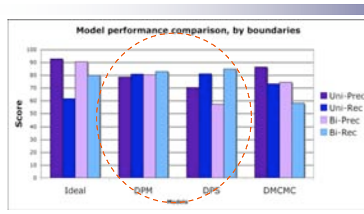
#correct / #found

Recall:

#found / #true

The DMCMC learner, like the Ideal learner, tends to undersegment.

Results: under vs. oversegmentation



Precision:

#correct / #found

Recall:

#found / #true

The DPM and DPS learners, however, tend to oversegment.

Results: interim summary

- While no online learners outperform the best ideal learner on all measures, all perform better on realistic child-directed speech data than a syllable transitional probability learner, which achieves a token F score of 29.9 (Gambell & Yang 2006).
- While assuming words are predictive units (bigram model) significantly helped the ideal learner, this assumption may not be as useful to an online learner (depending on how memory limitations are implemented).

Results: interim summary

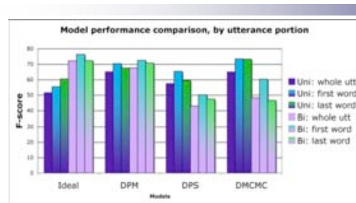
- The tendency to undersegment the corpus also depends on how memory limitations are implemented. Undersegmentation may match children's performance better than oversegmentation (Peters 1983).
- The lower the decay rate in the DMCMC learner, the more the learner tends to undersegment. (Ask for details!)

Results: Exploring different performance measures

- Some positions in the utterance are more easily segmented by infants, such as the first and last word of the utterance (Seidl & Johnson 2006).
  - The first and last word are less ambiguous (one boundary known) (first = last > whole utterance)
  - Memory effects & prosodic prominence make the last word easier (last > first, whole utterance)
  - The first/last word are more regular, due to syntactic properties (first, last > whole utterance)

```
look theres a boy with his hat
and a doggie
you want to look at this
Look at this
```

Results: Exploring different performance measures

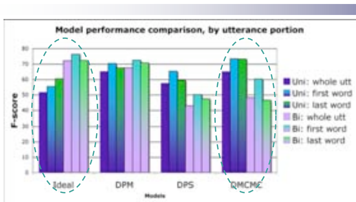


Results from 2nd half of corpus

Unigrams vs. Bigrams, Token F-scores

whole utterance
first word
last word

Results: Exploring different performance measures



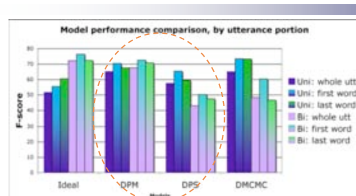
Unigrams vs. Bigrams, Token F-scores

whole utterance
first word
last word

The Ideal unigram learner performs better on the first and last words in the utterance, while the bigram learner only improves for the first words. The DMCMC follows this trend.

Unigram: first ≤ last > whole utterance
Bigram: first > last, whole utterance

Results: Exploring different performance measures



Unigrams vs. Bigrams, Token F-scores

whole utterance
first word
last word

The DPM and DPS learners always improve on the first and last words, irrespective of n-gram model. The first word tends to improve more than the last word.

Unigram/Bigram: first > last > whole utterance

Summary: Online Learners

- Simple intuitions about human cognition (e.g. memory limitations) can be translated in multiple ways
  - processing utterances incrementally
  - keeping a single lexicon hypothesis in memory
  - implementing recency effects
- Learning biases/assumptions that are helpful in an ideal learner may hinder a learner with processing constraints. However, constrained learners can still use statistical regularity available in the data.
- Statistical learning doesn't have to be perfect to reflect acquisition: online statistical learning may provide a lexicon reliable enough for children to learn language-dependent strategies from.

The End & Thank You!

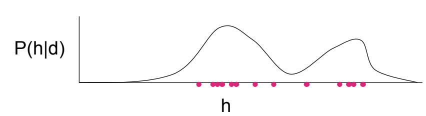
Special thanks to…
Tom Griffiths
Michael Frank
the Computational Models of Language Learning Seminar at UCI

This work was supported by NSF grant BCS-0843896 to LP.

Search algorithm comparison

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.



- Ideal: A batch algorithm

  vs. DMCMC: incremental algorithm that uses the same sampling equation

Gibbs sampler

- Compares pairs of hypotheses differing by a single word boundary:

```
whats.that            whats.that
the.doggie            the.dog.gie
yeah                  yeah
wheres.the.doggie     wheres.the.doggie
…                     …
```

- Calculate the probabilities of the words that differ, given current analysis of all other words.
- Sample a hypothesis according to the ratio of probabilities.

The unigram model

Assumes word w_i is generated as follows:

1. Is w_i a novel lexical item?

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

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

Fewer word types = Higher probability

The unigram 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

Notes

- Distribution over words is a Dirichlet Process (DP) with concentration parameter α and base distribution P_0:

$$P(w_i = w \mid w_1 \ldots w_{i-1}) = \frac{n_w + \alpha P_0(w)}{i - 1 + \alpha}$$

- Also (nearly) equivalent to Anderson's (1990) Rational Model of Categorization.

Bigram model

Assume word w_i is generated as follows:

1. Is (w_{i-1}, w_i) a novel bigram?

$$P(yes) = \frac{\beta}{n_{w_{i-1}} + \beta} \qquad P(no) = \frac{n_{w_{i-1}}}{n_{w_{i-1}} + \beta}$$

2. If novel, generate w_i using unigram model (almost).

   If not, choose lexical identity of w_i from words previously occurring after w_{i-1}.

$$P(w_i = w \mid w_{i-1} = w') = \frac{n_{(w',w)}}{n_{w'}}$$

Notes

- Bigram model is a hierarchical Dirichlet process (Teh et al., 2005):

$$P(w_i = w \mid w_{i-1} = w', w_1 \ldots w_{i-2}) = \frac{n_{(w',w)} + \beta P_1(w)}{i - 1 + \beta}$$

$$P_1(w_i = w \mid w_1 \ldots w_{i-1}) = \frac{b_w + \alpha P_0(w)}{b + \alpha}$$

Results: Exploring decay rates in DMCMC

Unigram learners, *s* = 10000

	Word Tokens Prec	Word Tokens Rec	Boundaries Prec	Boundaries Rec	Lexicon Prec	Lexicon Rec
***d*=2**	23.8	36.7	45.2	**80.0**	14.9	13.6
***d*=1.5**	59.9	53.4	75.4	68.7	30.2	38.7
***d*=1**	**69.1**	**61.6**	86.4	73.2	51.1	54.1
***d*=0.75**	58.7	61.0	86.2	72.5	**54.0**	55.9
***d*=0.5**	64.0	53.0	87.7	66.3	51.3	55.6
***d*=0.25**	60.6	47.4	**88.3**	61.0	48.0	**57.4**

- Decay rate 1 has best performance by tokens.
- Undersegmentation occurs more as decay rate decreases.
- Lexicon recall increases as decay rate decreases, and is generally higher than lexicon precision.

Results: Exploring decay rates in DMCMC

Bigram learners, *s* = 15000

	Word Tokens Prec	Word Tokens Rec	Boundaries Prec	Boundaries Rec	Lexicon Prec	Lexicon Rec
***d*=2**	40.1	38.9	61.6	**59.0**	15.5	38.5
***d*=1.5**	45.0	41.3	66.9	**59.0**	16.6	38.0
***d*=1**	54.0	45.7	75.4	**59.0**	**19.3**	**42.7**
***d*=0.75**	51.0	43.6	74.1	58.8	18.2	40.5
***d*=0.5**	**54.9**	**45.9**	**76.8**	58.8	17.5	38.5
***d*=0.25**	53.2	43.7	76.3	57.0	18.2	41.3

- Decay rate 0.5 has the best performance by tokens.
- Undersegmentation still occurs more as decay rate decreases.
- Lexicon precision suffers significantly, compared to the unigram learners.