From sounds to words:
Bayesian modelling of early language acquisition

Sharon Goldwater

The problem

```
+    “Look at the doggie”
```

A multi-layered problem

- Phonetics:
  - /n/
- Word segmentation:
  - see the doggie
- Phonotactics:
  - kell > shrem > vlep
- Phonology:
  - Two wug[z] vs. two blick[s]
- Morphology:
  - he’s foozing => he foozed
- Syntax:
  - ‘Look at the big dog’ vs. ‘At the big dog look’
- Semantics:
  - big (dog) ≠ big (house)

Language learning as induction

```
Input
(specific linguistic observations)
```
```
Grammar/lexicon
(abstract internal representation)
```
```
Output
(specific linguistic productions)
```

Sources of constraints

- Innate constraints:
  - Domain-general: memory, perception, reasoning, categorization.
  - Domain-specific: inventory of syntactic categories, rules, principles, parameters, etc.
- Previously acquired knowledge (bootstrapping):
  - She lumped heavily into the room.

- How do these interact with each other and the input?

Modeling approach

- Questions can be addressed within a Bayesian framework – a structured probabilistic approach.
  - Probabilistic: learner can exploit partial or uncertain information to help solve the bootstrapping problem.
  - Structured: models explicitly define representations, biases (constraints), and use of information.
Bayesian modeling

- An ideal observer approach.
  - What is the optimal solution to the induction problem, given particular assumptions about representation and available information?
  - In what ways might humans differ from this ideal learner, and why?

Outline

1. Introduction
2. Word segmentation, computational model and theoretical results
   (joint work with Tom Griffiths and Mark Johnson)
3. Modeling experimental data
   (joint work with Mike Frank, Vikash Mansinghka, Tom Griffiths, and Josh Tenenbaum)

Word segmentation

Input: continuous speech
Output: segmented word tokens

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.

Statistical segmentation

  - $P(y_t | y_{t-1})$ is often lower at word boundaries.
- What do TPs have to say about words?
  1. A word is a unit whose beginning predicts its end, but it does not predict other words.
  2. A word is a unit whose beginning predicts its end, and it also predicts future words.

Focusing on words

- Most previous work assumes words are statistically independent.
  - Experimental work: Saffran et al. (1996), many others.
  - Computational work: Brent (1999).
- What about words predicting other words?
Questions

- If a learner assumes that words are independent units, what is learned (from more realistic input)?
- What if the learner assumes that words are units that help predict other units?

Approach: use a Bayesian ideal observer model to examine the consequences of making these different assumptions. What kinds of words are learned?

Two kinds of models

- Unigram model: words are independent.
  - Generate a sentence by generating each word independently.
  - Unigram: look \(.1\) that \(.2\) at \(.4\) ...
  - Bigram: look \(.1\) at \(.2\) that \(.2\) at \(.4\) ...

Bayesian learning

- The Bayesian learner seeks to identify an explanatory linguistic hypothesis that:
  - accounts for the observed data.
  - conforms to prior expectations.
  - Focus is on the goal of computation, not the procedure (algorithm) used to achieve the goal.

Bayesian segmentation

- In the domain of segmentation, we have:
  - Data: unsegmented corpus (transcriptions).
  - Hypotheses: sequences of word tokens.
  - Optimal solution is the segmentation with highest prior probability.

Data:

- look at the doggie
- see the doggie
- she looks so friendly
- I like pizza
- what about you

Hypotheses:

- look at the doggie
- see the doggie
- she looks so friendly
- I like pizza
- what about you

$P(d|h) = 1$ if concatenating words forms corpus, $P(d|h) = 0$ otherwise.

Encodes assumptions of learner.
Brent (1999)

- Describes a Bayesian unigram model for segmentation.
  - Prior favors solutions with fewer words, shorter words.
- Problems with Brent’s system:
  - Learning algorithm is approximate (non-optimal).
  - Difficult to extend to incorporate bigram info.

Bayesian model

Assumes word $w_i$ is generated as follows:

1. Is $w_i$ a novel lexical item?

$$P(\text{yes}) = \frac{\alpha}{n + \alpha}$$  \hspace{1cm} \text{Fewer word types = Higher probability}

$$P(\text{no}) = \frac{n}{n + \alpha}$$  \hspace{1cm} \text{Power law = Higher probability}

Bayesian model: simulations

- Same corpus as Brent:
  - 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:

  `yuwanttus1d6b0k1UK*2e6b7wI2h1h1ht4nd6d06p1yuwanttul1k4tDIe...`

Results

- Example segmentation:

  `yuwant to see the book
  look theress aboy with his hat and doggie
  you want to look at this
  look at this
  have a drink
  okay now
  what this
  what that
  what is it
  look can you take it out
  ...`
Results

- Proposed boundaries are more accurate than Brent’s, but fewer proposals are made.

<table>
<thead>
<tr>
<th>Boundary Precision</th>
<th>Boundary Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brent</td>
<td>.80</td>
</tr>
<tr>
<td>GGJ</td>
<td>.92</td>
</tr>
</tbody>
</table>

Result: word tokens are less accurate.

<table>
<thead>
<tr>
<th>Token F-score</th>
<th>F-score: an average of precision and recall.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brent</td>
<td>.66</td>
</tr>
<tr>
<td>GGJ</td>
<td>.54</td>
</tr>
</tbody>
</table>

What happened?

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

\[
P(\text{that}) = .024 \quad P(\text{that} | \text{whats}) = .46 \quad P(\text{that} | \text{to}) = .0019
\]

- positing amalgams allows the model to capture word-to-word dependencies.

What about other unigram models?

- Brent’s learning algorithm is insufficient to identify the optimal segmentation.
  - Our solution has higher probability under his model than his own solution does.
  - On randomly permuted corpus, our system achieves 96% accuracy; Brent gets 81%.

- Formal analysis shows undersegmentation is the optimal solution for any (reasonable) unigram model.

Bigram model

Assume word \( w_i \) is generated as follows:

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

\[
P(\text{yes}) = \frac{\beta}{n_{w_i} + \beta} \quad P(\text{no}) = \frac{n_{w_i}}{n_w + \beta}
\]

2. If novel, generate \( w_i \) using unigram model (almost).
   - If not, choose lexical identity of \( w_i \) from words previously occurring after \( w_{i-1} \).

\[
P(w_i = w | w_{i-1} = w') = \frac{n_{w_i} | w'}{n_{w_{i-1}}}
\]

Results

- Example segmentation:

  you want to see the book
  look there’s a boy with his hat
  and a doggie
  you want to look at this
  look at this
  have a drink
  okay now
  what’s this
  what is it
  look can you take it out
  …

- Compared to unigram model, more boundaries are proposed, with little loss in accuracy.

<table>
<thead>
<tr>
<th>Boundary Precision</th>
<th>Boundary Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>GGJ (unigram)</td>
<td>.92</td>
</tr>
<tr>
<td>GGJ (bigram)</td>
<td>.90</td>
</tr>
</tbody>
</table>

- Accuracy is higher than previous models:

<table>
<thead>
<tr>
<th>Token F-score</th>
<th>Type F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brent (unigram)</td>
<td>.68</td>
</tr>
<tr>
<td>GGJ (bigram)</td>
<td>.72</td>
</tr>
</tbody>
</table>
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.
  - Units tend to come in predictable patterns.

Remaining questions

- Is unigram segmentation sufficient to start bootstrapping other cues (e.g., stress)?
- How prevalent are multi-word chunks in infant vocabulary?
- Are humans able to segment based on bigram statistics?
- Is there any evidence that human performance is consistent with Bayesian predictions?

Outline

1. Introduction
2. Word segmentation, computational model and theoretical results
   (joint work with Tom Griffiths and Mark Johnson)
3. Modeling experimental data
   (joint work with Mike Frank, Vikash Mansinghka, Tom Griffiths, and Josh Tenenbaum)

Testing model predictions

- Saffran-style experiment using multiple utterances.
  - Synthesize stimuli with 500ms pauses between utterances.
  - Training: adult subjects listen to corpus of utterances.
  - Testing: 2AFC between words and part-word distractors

Experiment 1: utterance length

- Vary the number of words per utterance.

<table>
<thead>
<tr>
<th>#vocab</th>
<th># wds/utt</th>
<th># utts</th>
<th>tot # wds</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>1200</td>
<td>1200</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>600</td>
<td>1200</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>300</td>
<td>1200</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>200</td>
<td>1200</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>150</td>
<td>1200</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>100</td>
<td>1200</td>
</tr>
</tbody>
</table>

Experiment 2: exposure time

- Vary the number of utterances heard in training.

<table>
<thead>
<tr>
<th>#vocab</th>
<th># wds/utt</th>
<th># utts</th>
<th>tot # wds</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>4</td>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>25</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>75</td>
<td>300</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>150</td>
<td>600</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>225</td>
<td>900</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>300</td>
<td>1200</td>
</tr>
</tbody>
</table>
Experiment 3: vocabulary size

- Vary the number of lexical items.

<table>
<thead>
<tr>
<th>vocab</th>
<th># wds/utt</th>
<th># utts</th>
<th>tot # wds</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
<td>150</td>
<td>600</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>150</td>
<td>600</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>150</td>
<td>600</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>150</td>
<td>600</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>150</td>
<td>600</td>
</tr>
</tbody>
</table>

Human results: utterance length

Human results: exposure time

Human results: vocabulary size

Model comparison

- Evaluated six different models.
- Each model trained and tested on same stimuli as humans.
- For testing, produce a score $s(w)$ for each item in choice pair and use Luce choice rule:
\[
P(w_i) = \frac{s(w_i)}{s(w_i) + s(w_j)}
\]
- Calculate correlation coefficients between each model’s results and the human data.

Models used

- Several variations on transitional probabilities (TP)
  - $s(w) =$ minimum TP in $w$.
- Swingley (2005)
  - Builds lexicon using local statistic and frequency thresholds.
  - $s(w) =$ max threshold at which $w$ appears in lexicon.
- PARSER (Perruchet and Vintner, 1998)
  - Incorporates principles of lexical competition and memory decay.
  - $s(w) =$ $P(w)$ as defined by model.
- Bayesian model
  - $s(w) =$ $P(w)$ as defined by model.
Results: utterance length

<table>
<thead>
<tr>
<th></th>
<th>Transitional probability</th>
<th>Bayesian model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance length (words)</td>
<td><img src="image" alt="Graph" /> r = 0.84</td>
<td><img src="image" alt="Graph" /> r = 0.93</td>
</tr>
</tbody>
</table>

Swingley (2005)  PARSER

Results: exposure time

<table>
<thead>
<tr>
<th></th>
<th>Transitional probability</th>
<th>Bayesian model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of exposure (tokens)</td>
<td><img src="image" alt="Graph" /> r = 0.43</td>
<td><img src="image" alt="Graph" /> r = 0.91</td>
</tr>
</tbody>
</table>

Swingley (2005)  PARSER

Summary: Experiments 1 and 2

- For humans, learning to segment is more difficult
  - when utterances contain more words.
  - when less data is available.
- Only Bayesian model captures both effects:
  - Success is due to accumulation of evidence for best hypothesis, moderated by competition with other hypotheses.

<table>
<thead>
<tr>
<th></th>
<th>TPs</th>
<th>Swidt</th>
<th>PARSER</th>
<th>Bayes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance length</td>
<td>✓</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Exposure</td>
<td>x</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Model results: vocabulary size

<table>
<thead>
<tr>
<th></th>
<th>Transitional probability</th>
<th>Bayesian model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary size (types)</td>
<td><img src="image" alt="Graph" /> r = -0.99</td>
<td><img src="image" alt="Graph" /> r = -0.99</td>
</tr>
</tbody>
</table>

Swingley (2005)  PARSER

What's going wrong?

- TPs: smaller vocab => TPs across words are higher.
- Bayes: smaller vocab => Incorrect solutions have relatively small vocabularies with many frequent "words".

- With perfect memory, stronger statistical cues of larger vocabulary outweigh increased storage needs.

Memory limitations

- Modified Bayesian model has limited memory for data and generalizations.
  - Online learning algorithm processes one utterance at a time, one pass through data.
  - Random decay of items in lexicon.
- Learner is no longer guaranteed to find optimal solution.
Results: memory-limited learner

- Good fit to all three experiments:
  - Simulating limited memory in TP also improves results but not as much.

Summary

- Humans behave like ideal learners in some cases.
  - Longer utterances are harder – competition.
  - Shorter exposure is harder – less evidence.
- Humans are unlike ideal learners in other cases.
  - Larger vocabulary is harder for humans, easier for model.
- Memory-limited learner captures human behavior in all three experiments.

Conclusions

- Bayesian modeling provides a framework for investigating the relationship between linguistic input and the learner's representations and constraints.
- Work on word segmentation suggests
  - General constraints may be sufficient for this task.
  - Word-based (not boundary-based) representations are important for word segmentation.
  - Humans behave like ideal learners in some respects.
  - Accounting for limited memory is important.

Further details and extensions


Online algorithms:

Noisy input data:

Targets vs. distractors
Inference

- We use a Gibbs sampler that compares pairs of hypotheses differing by a single word boundary:

<table>
<thead>
<tr>
<th>what</th>
<th>the</th>
<th>doggie</th>
</tr>
</thead>
<tbody>
<tr>
<td>yeah</td>
<td>where</td>
<td>the</td>
</tr>
</tbody>
</table>

- Calculate the probabilities of the words that differ, given current analysis of all other words.

- Sample a hypothesis according to the ratio of probabilities.

Incremental Sampling

For each utterance:
- Sample a segmentation from the posterior distribution given the current lexicon.
- Add counts of segmented words to lexicon.

- Online algorithm
- Limits memory for corpus data

(Particle filter: more particles ⇔ more memory)