Synergies in learning words and their referents

Mark Johnson\textsuperscript{1}, Katherine Demuth\textsuperscript{1}, Michael Frank\textsuperscript{2} and Bevan Jones\textsuperscript{3}

\textsuperscript{1}Macquarie University
\textsuperscript{2}Stanford University
\textsuperscript{3}University of Edinburgh

NIPS 2010
Two hypotheses about language acquisition

1. Pre-programmed *staged acquisition* of linguistic components
   - “*Semantic bootstrapping*”: semantics is learnt first, and used to predict syntax (Pinker 1984)
   - “*Syntactic bootstrapping*”: syntax is learnt first, and used to predict semantics (Gleitman 1991)
   - Conventional view of *lexical acquisition*, e.g., Kuhl (2004)
     - child first learns the phoneme inventory, which it then uses to learn
     - phonotactic cues for word segmentation, which are used to learn
     - phonological forms of words in the lexicon, ...

2. *Interactive acquisition* of all linguistic components together
   - corresponds to *joint inference* for all components of language
   - stages in language acquisition might be due to:
     - child’s input may contain more information about some components
     - some components of language may be learnable with less data
Synergies: an advantage of interactive learning

• An interactive learner can take advantage of synergies in acquisition
  ▶ partial knowledge of component A provides information about component B
  ▶ partial knowledge of component B provides information about component A

• A staged learner can only take advantage of one of these dependencies

• An interactive learner can benefit from a positive feedback cycle between A and B

• This paper investigates whether there are synergies in learning how to segment words and learning the referents of words
Prior work: mapping words to referents

• Input to learner:
  ▶ word sequence: *Is that the pig?*
  ▶ objects in nonlinguistic context: DOG, PIG

• Learning objectives:
  ▶ identify utterance topic: PIG
  ▶ identify word-topic mapping: *pig* $\mapsto$ PIG
Frank et al (2009) “topic models” as PCFGs

• Prefix each sentence with possible topic marker, e.g., PIG|DOG
• PCFG rules designed to choose a topic from possible topic marker and propagate it through sentence
• Each word is either generated from sentence topic or null topic Ø
• Simple grammar modification requires at most one topical word per sentence
• Bayesian inference for PCFG rules and trees corresponds to Bayesian inference for word and sentence topics using topic model (Johnson 2010)
Prior work: segmenting words in speech

• Running speech does not contain “pauses” between words ⇒ child needs to learn how to segment utterances into words
• Elman (1990) and Brent et al (1996) studied segmentation using an artificial corpus
  ▶ child-directed utterance: *Is that the pig?*
  ▶ broad phonemic representation: /ɪz ðæt ðə pɪɡ/
  ▶ input to learner: \[ N \_ N \_ N \_ N \_ \]
  [i z] [d æ t] [d e] [p i g]
• Learner’s task is to identify which potential boundaries correspond to word boundaries
Brent (1999) unigram model as adaptor grammar

- Adaptor grammars (AGs) are CFGs in which a subset of nonterminals are adapted
  - AGs learn probability of entire subtrees of adapted nonterminals (Johnson et al 2007)
  - AGs are hierarchical Dirichlet or Pitman-Yor Processes
  - Prob. of adapted subtree $\propto$ number of times tree was previously generated $+$ $\alpha \times$ PCFG prob. of generating tree

- AG for unigram word segmentation:
  
  Words $\rightarrow$ Word $|$ Word Words
  
  Word $\rightarrow$ Phons
  
  Phons $\rightarrow$ Phon $|$ Phon Phons

  (Adapted nonterminals indicated by underlining)
Prior work: Collocation AG (Johnson 2008)

- Unigram model doesn’t capture *interword dependencies*
  ⇒ tends to *undersegment* (e.g., *ɪz ðæt ðəpɪg*)
- Collocation model “explains away” some interword dependencies
  ⇒ more accurate word segmentation

Sentence → Colloc+
Colloc → Word+
Word → Phon+

- Kleene “+” abbreviates right-branching rules
- Unadapted internal nodes suppressed in trees
AGs for joint segmentation and referent-mapping

- Easy to combine topic-model PCFG with word segmentation AGs
- Input consists of unsegmented phonemic forms prefixed with possible topics:

\[
\text{PIG}|\text{DOG} \ i \ z \ \delta \ \alpha \ t \ \delta \ \partial \ p \ i \ g
\]

- E.g., combination of *Frank “topic model”* and *unigram segmentation model*
  - equivalent to Jones et al (2010)

- Easy to define *other combinations of topic models and segmentation models*
Collocation topic model AG

- Collocations are either “topical” or not
- Easy to modify this grammar so
  - at most one topical word per sentence, or
  - at most one topical word per topical collocation
Experimental set-up

- Input consists of unsegmented phonemic forms prefixed with possible topics:
  
  \[ \text{PIG|DOG I z ð æ t ð ñ p ñ g} \]

  - Child-directed speech corpus collected by Fernald et al (1993)
  - Objects in visual context annotated by Frank et al (2009)

- Bayesian inference for AGs using MCMC (Johnson et al 2009)
  - Uniform prior on PYP $a$ parameter
  - “Sparse” Gamma($100, 0.01$) on PYP $b$ parameter

- For each grammar we ran 8 MCMC chains for 5,000 iterations
  - collected word segmentation and topic assignments at every 10th iteration during last 2,500 iterations
    $\Rightarrow$ 2,000 sample analyses per sentence
  - computed and evaluated the modal (i.e., most frequent) sample analysis of each sentence
Does non-linguistic context help segmentation?

<table>
<thead>
<tr>
<th>Model segmentation</th>
<th>topics</th>
<th>word segmentation token f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigram</td>
<td>not used</td>
<td>0.533</td>
</tr>
<tr>
<td>unigram</td>
<td>any number</td>
<td>0.537</td>
</tr>
<tr>
<td>unigram</td>
<td>one per sentence</td>
<td>0.547</td>
</tr>
<tr>
<td>collocation</td>
<td>not used</td>
<td>0.695</td>
</tr>
<tr>
<td>collocation</td>
<td>any number</td>
<td>0.726</td>
</tr>
<tr>
<td>collocation</td>
<td>one per sentence</td>
<td>0.719</td>
</tr>
<tr>
<td>collocation</td>
<td>one per collocation</td>
<td>0.750</td>
</tr>
</tbody>
</table>

- Not much improvement with unigram model
  - consistent with results from Jones et al (2010)
- Larger improvement with collocation model
  - most gain with *one topical word per topical collocation* (this constraint cannot be imposed on unigram model)
Does better segmentation help topic identification?

- Task: identify object (if any) *this sentence* is about

<table>
<thead>
<tr>
<th>Model segmentation</th>
<th>topics</th>
<th>sentence referent accuracy</th>
<th>f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigram</td>
<td>not used</td>
<td>0.709</td>
<td>0</td>
</tr>
<tr>
<td>unigram</td>
<td>any number</td>
<td>0.702</td>
<td>0.355</td>
</tr>
<tr>
<td>unigram</td>
<td>one per sentence</td>
<td>0.503</td>
<td>0.495</td>
</tr>
<tr>
<td>collocation</td>
<td>not used</td>
<td>0.709</td>
<td>0</td>
</tr>
<tr>
<td>collocation</td>
<td>any number</td>
<td>0.728</td>
<td>0.280</td>
</tr>
<tr>
<td>collocation</td>
<td>one per sentence</td>
<td>0.440</td>
<td>0.493</td>
</tr>
<tr>
<td>collocation</td>
<td>one per collocation</td>
<td>0.839</td>
<td>0.747</td>
</tr>
</tbody>
</table>

- The collocation grammar with *one topical word per topical collocation* is the only model clearly better than baseline
Does better segmentation help topic identification?

- Task: identify *head nouns* of NPs referring to topical objects (e.g. *pig* → PIG in input PIG | DOG I z d æ t ð æ p I g)

<table>
<thead>
<tr>
<th>Model segmentation</th>
<th>topics</th>
<th>topical word f-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>unigram</td>
<td>not used</td>
<td>0</td>
</tr>
<tr>
<td>unigram</td>
<td>any number</td>
<td>0.149</td>
</tr>
<tr>
<td>unigram</td>
<td>one per sentence</td>
<td>0.147</td>
</tr>
<tr>
<td>collocation</td>
<td>not used</td>
<td>0</td>
</tr>
<tr>
<td>collocation</td>
<td>any number</td>
<td>0.220</td>
</tr>
<tr>
<td>collocation</td>
<td>one per sentence</td>
<td>0.321</td>
</tr>
<tr>
<td>collocation</td>
<td>one per collocation</td>
<td><strong>0.636</strong></td>
</tr>
</tbody>
</table>

- The collocation grammar with one topical word per topical collocation is best at identifying head nouns of referring NPs
Conclusions and future work

- Adaptor Grammars can express a variety of useful HDP models
  - generic AG inference code makes it easy to explore models
- There seem to be synergies a learner could exploit when learning word segmentation and word-object mappings
  - incorporating word-topic mapping improves segmentation accuracy (at least with collocation grammars)
  - improving segmentation accuracy improves topic detection and acquisition of topical words
- Caveat: results seem to depend on details of model
- Future work:
  - extend expressive power of AGs (e.g., phonology, syntax)
  - richer data (e.g., more non-linguistic context)
  - more realistic data (e.g., phonological variation)