Synergies in learning words and their referents

Mark Johnson[1], Katherine Demuth[1], Michael Frank[2] and Bevan Jones[3]

1: Macquarie University, 2: Stanford University, 3: University of Edinburgh

Abstract

- *Question: is information from the non-linguistic context useful in learning to identify words?*
- *Previous work has identified Bayesian methods for:*
  - *identifying words in an unsegmented stream of phonemes (Goldwater et al 2009)*
  - *the mapping from words to objects they refer to (Frank et al 2009)*
- *Both of these models can be expressed as adaptor grammars (Johnson et al 2007)*
  - *adaptor grammars specify Hierarchical Dirichlet Processes over trees generated by CFGs*
  - *"rich get richer" ⇒ frequently appearing subtrees are more likely to be reused*
- *We show how to construct adaptor grammars that perform word segmentation and map the words they learn to objects*
- *The non-linguistic context permits our "one topic per collocation" adaptor grammar to learn words more accurately than corresponding adaptor grammars that don't use non-linguistic context.*

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

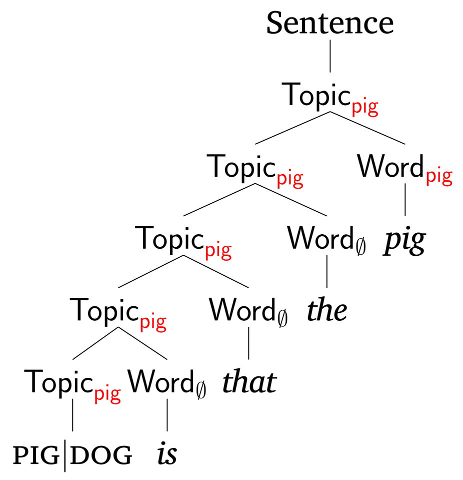
Synergies: an advantage of interactive learning

- 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* ↦ PIG

Frank et al (2009) "topic models" as PCFGs

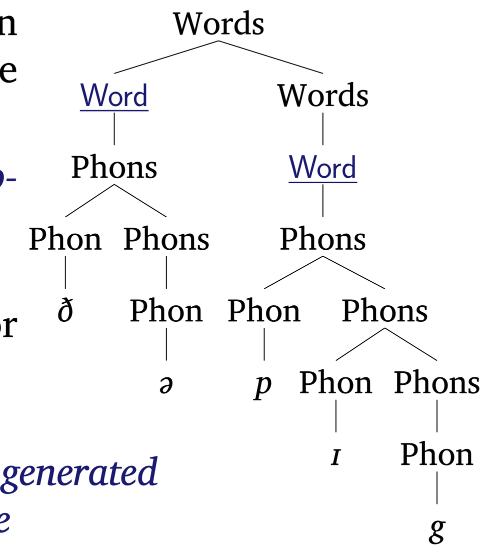


- Prefix each sentence with *possible topic marker*, e.g., PIG|DOG
- PCFG rules designed to *choose a topic* from possible topic marker and *propagate it through sentence*
- Each word is either generated from sentence topic or null topic ∅
- Simple grammar modification requires *at most one topical word 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 pɪg*
  - input to learner: ▵ɪ▵z▵ð▵æ▵t▵ð▵ə▵p▵ɪ▵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 ∝ *number of times tree was previously generated* $+ \alpha \times$ *PCFG prob. of generating tree*
- AG for *unigram word segmentation*:

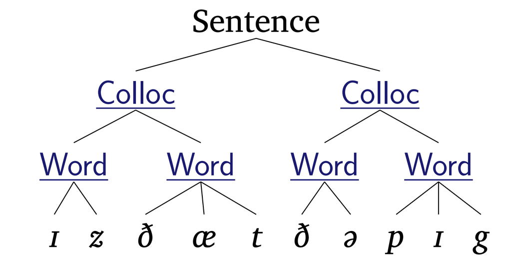
Words → Word | Word Words
<u>Word</u> → Phons
Phons → Phon | Phon Phons

(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 → <u>Colloc</u>$^+$
<u>Colloc</u> → <u>Word</u>$^+$
<u>Word</u> → Phon$^+$



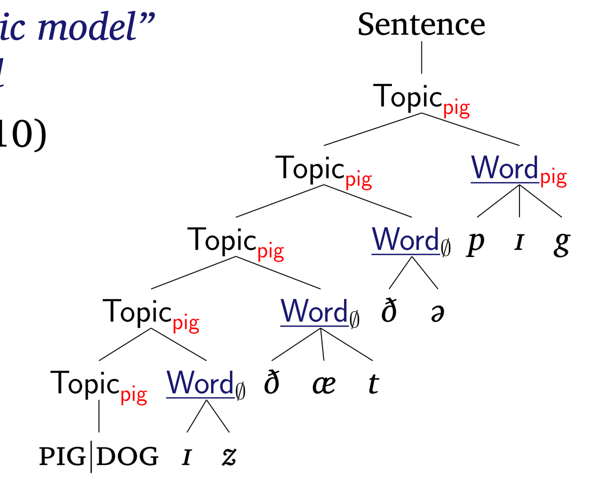
- Kleene "+" abbreviates right-branching rules
- Unadapted internal nodes suppressed in trees

AGs for joint segmentation and referent-mapping

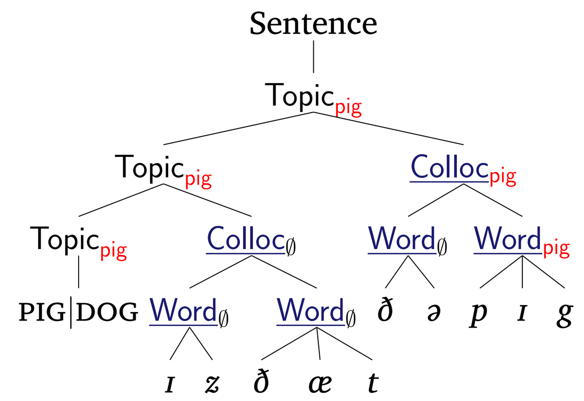
- Easy to combine topic-model PCFG with word segmentation AGs
- Input consists of unsegmented phonemic forms prefixed with possible topics:

  PIG|DOG *ɪ z ð æ t ð ə p ɪ 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:

  PIG|DOG *ɪ z ð æ t ð ə p ɪ 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" $\mathrm{Gamma}(100, 0.01)$ 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
    ⇒ 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?

Model		word segmentation
segmentation	topics	token f-score
unigram	not used	0.533
unigram	any number	0.537
unigram	one per sentence	0.547
collocation	not used	0.695
collocation	any number	0.726
collocation	one per sentence	0.719
collocation	one per collocation	**0.750**

- 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

Model		sentence referent	
segmentation	topics	accuracy	f-score
unigram	not used	0.709	0
unigram	any number	0.702	0.355
unigram	one per sentence	0.503	0.495
collocation	not used	0.709	0
collocation	any number	0.728	0.280
collocation	one per sentence	0.440	0.493
collocation	one per collocation	**0.839**	**0.747**

- 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 *ɪ z ð æ t ð ə p ɪ g*)

Model		topical word
segmentation	topics	f-score
unigram	not used	0
unigram	any number	0.149
unigram	one per sentence	0.147
collocation	not used	0
collocation	any number	0.220
collocation	one per sentence	0.321
collocation	one per collocation	**0.636**

- 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)