A log-linear model of language acquisition with multiple cues

Gabriel Doyle          Roger Levy
UC San Diego Linguistics
LSA 2011
mommyisntherenoweweateatyourapple
transition probabilities

phonotactics

allophonic variation

stress patterns

coarticulation

mommy isn't there now eat your apple

S W
no single sufficient cue

Vowel Categorization

Vallabha et al 2007, PNAS
Learning from Multiple Cues

• Linguistic problems can have multiple partially informative cues
• Need for models that learn to use cues jointly
The log-linear multi-cue model

- **General computational model** for learning structures from multiple cues

- **Specific implementation** in word segmentation using transition probabilities and stress patterns
Outline

• The Multiple-Cue Problem
• Case study: Word Segmentation
• Log-linear multiple-cue model
• Experimental testing
Case Study: Word Segmentation

• Transition probabilities
  – $p(B|A)$: probability that, having seen A, you’ll see B next

  \[ p(\text{key}|\text{mon}) = 1 \quad \text{p(hat|the)} = 1/2 \]

  Point to the monkey with the hat

  – Lower TP suggests separate words
  – 8 month old infants use TPs to segment artificial languages (Saffran et al 1996, a.o.)
Case Study: Word Segmentation

• **Stress patterns**
  
  – English has trochaic (Strong-Weak) bias
    
    **Double, double, toil and trouble; Fire burn and cauldron bubble**
  
  – 90% of content words start strong (Cutler & Carter 1987)
  
  – 7.5 month old English learners segment trochaic but not iambic words (Jusczyk et al 1999)
Existing segmentation models

• Single cue-type (phonemes)
  – Bayesian MDL models (Goldwater et al 2009)
  – PUDDLE (Monaghan & Christiansen 2010)

• Multi cue-type (phonemes & stress)
  – Connectionist (Christiansen et al 1998)
  – Algorithmic (Gambell & Yang 2006)
Why a log-linear model?

- Ideal learner model; other multi-cue models aren’t
- Effective in other linguistic tasks (Hayes & Wilson 2008, Poon et al 2009)
- More flexible than other models
  - new cues become new features
  - overlapping cues are easy to incorporate
Log-linear modelling

- Model learns a probability distribution

\[
p(W, S) = \frac{1}{Z} e^{\sum_j \lambda_j f_j(W, S)}
\]

- Feature functions \( f_j \) map \((W, S)\) pairs to real numbers

- “Learning” means finding good real number weights \( \lambda \) for features
Feature functions

• Transition probabilities
  – Bigram counts within words

• Stress templates
  – Stress “word” counts

• Lexical
  – Word counts

• MDL Prior
  – Lexicon length
“Normalizing” the probability

\[ p(W, S) = \frac{1}{Z} e^{\sum_j \lambda_j f_j(W, S)} \]

- Probabilities need to be normalized
- Usually divide by sum
- But this sum is intractable
Contrastive estimation

all possible corpora

observed corpus

contrast set
Contrastive estimation
(Smith & Eisner 2005)

• Contrast set as focused negatives
  – Want to put probability mass on grammatical outcomes
  – AND remove mass from ungrammaticals

• Good contrast sets can cause quicker convergence
Our contrast set

- Set of all corpora from transposing two syllables in observed corpus

Observed corpus \[\rightarrow\] mommy ate it

Ungrammatical contrasts

- mmymo ate it
- moate mmy it

“Grammatical” contrast \[\rightarrow\] mommy it ate

Note: not the only possible contrast set
Learning the weights $\lambda$

- Weights estimated using gradient ascent

$$\frac{\delta}{\delta \lambda_i} L(W^*) = E_{S|W^*}[f_i] - E_{S,W}[f_i] - \frac{1}{\sigma^2} (\lambda_i - \mu_i)$$

- Expected feature value on observed corpus
- Expected feature value on contrast set
- Prior

- Weight increases when feature appears in observed, decreases when it appears in contrast

- Prior pulls weight toward initial bias $\mu_i$
Experimental Questions

• Verification: Does it learn the stress biases that children exhibit?
  Training on child-directed English

• Application: Can these biases explain age effects in word segmentation?
  Testing on artificial language
Thiessen & Saffran 2003

- Synthesized bisyllabic language, either all SW or all WS
- 7 & 9 month olds, learning English
- Preferential looking after exposure
- Words & part words in opposition
Both ages segment by TPs & stress bias

7 mos: dobi > bibu
9 mos: dobi > bibu

7 mos seg by TPs
9 mos seg against TPs & with stress bias
Experimental Design

- Train on English child-directed speech
  - 1638 words of Pearl-Brent database
  - 266 SW, 35 WS; 80% monosyllabic
  - Stress determined by CMU Pron Dict
  - Utterance & syllable boundaries included, non-utterance word boundaries not given
  - no prior knowledge given
Weights learned from child-directed English

\[ \text{Mean } \lambda_{SW} - \lambda_{WS} = 0.262 \pm 0.119 \ [p < 0.001] \]
Age effects

- Idea: older infants have stronger confidence in language parameters
- Strength of learned priors increases to simulate increased linguistic experience

$$\frac{\delta}{\delta \lambda_i} L(W^*) = E_{S|W^*}[f_i] - E_{S,W}[f_i] - \frac{1}{\sigma^2}(\lambda_i - \mu_i)$$

prior strength

prior value
Age effects

7 months

Looking time

Word
Partword

SW
WS

9 months

-looking time

Word
Partword

SW
WS

“Young” model

Word score

SW
WS

“Old” model

Word score

SW
WS
Conclusions

- Model learns stress bias from unsegmented data
- Model shows similar behavioral change to infants learning a language
- Behavioral change can result strictly from exposure, not a change in the segmentation method
Future Extensions

- Expand set of cues (e.g., phonotactics)
- Additional experimental applications
- Move into other linguistic problems
Thank you!

gdoyle@ling.ucsd.edu