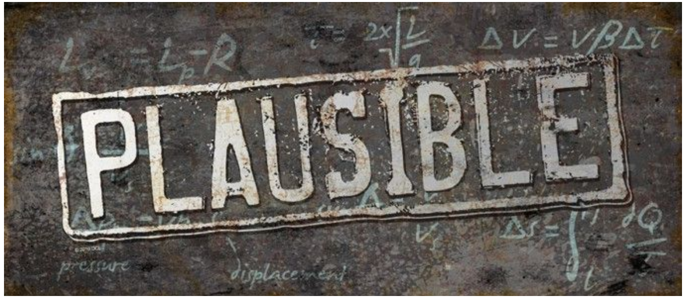
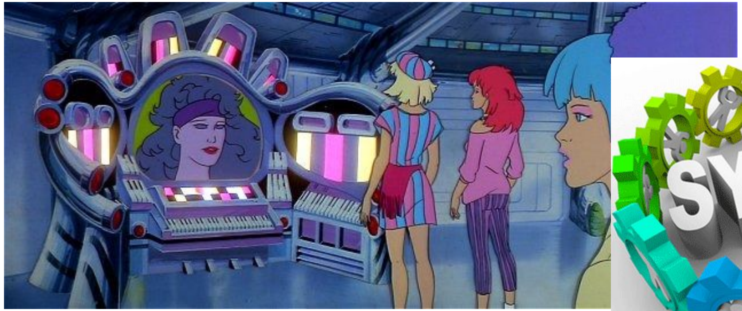
Cognitively 

Stephen Bennett

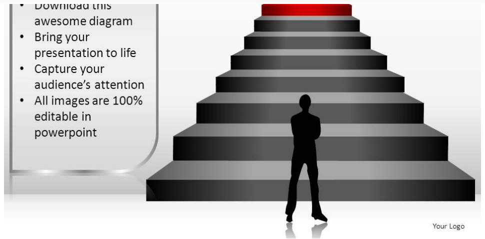
Overview

- Types of constraints
- Explicit application of these constraints

Overview



- Types of constraints
- Explicit application of these constraints
- Synergies from constraining several things at once



Algorithmic Constrains Computational

- Hypothesis space: Established theory and modern modeling inform which hypothesis spaces are possible
    - Modeling has to explicitly define the hypothesis space and therefore make large assumptions about what universal assumptions children use in learning language

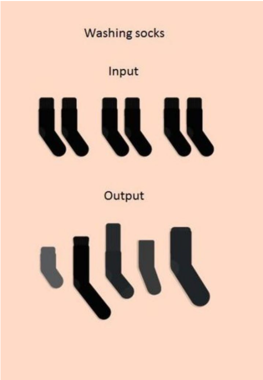
- Inference: Computational and Algorithmic diverge here

    - Computational takes the optimal inference given the problem & data
    - Algorithmic takes the most “appropriate” inference based on our knowledge of humans

Input/Output Constraints

- Input: Match as closely as possible with the input representation of the child

- Output: Instead of comparing against adult-level knowledge, compare the output against child-level trends.
    - Endgoal: Produce a representation of word segmentation in-line with children's representations based on background literature



Our Bayesian Strategy

$$P(h|d) \propto P(d|h)P(h)$$

- Ideal Learner-type model (Goldwater et al., 2009)
- Infer the simplest plausible lexicon
  - Smaller lexicon, shorter words in the lexicon
- Two implementations:
  - Unigram: No relationship between types of words that occur in sequence
  - Bigram: The preceding word informs which word is likely to follow

Our Bayesian Strategy: Model #1

- Generative Rule: $P(w_i|w_1,\ldots,w_{i-1}) = \dfrac{n_{i-1}(w_i) + \alpha P_0(w_i)}{i-1+\alpha}$
- $n_{i-1}(w_i)$ is the number of times w_i occurs in the previous i-1 words
- α relates the likelihood of a novel word (and is free to vary, so we have to consider its prior, which is presumably concentrated around low values)
- P_0 is the specifications for the composition of a novel word - how likely it is to be composed of certain phonemes or syllables
    - Enforces parsimony
    - Infers rules about the language (assuming P_0 is free to vary, which was not 100% clear)

$$P_0 = P(w = x_1,\ldots,x_m) = \prod_j P(x_j)$$

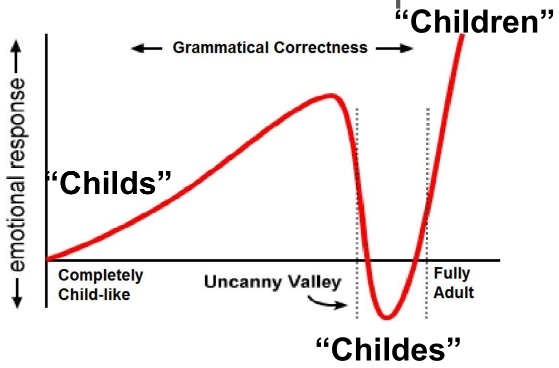
Our Bayesian Strategy: Model #2

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

- P_1 is equivalent to the equation on the previous slide: $P(w_i)$
- $n_{i-1}(w',w_i)$ is the number of times the bigram w',w_i occurs in the previous i words
- $n_{i-2}(w')$ is the number of times w’ occurs
- β is a free parameter that controls how strong of a bias towards few bigrams the model has

Input

- Child-directed speech - 9 months or younger from Pearl-Brent derived corpus of CHILDES
- Infants use a mix of inputs:
    - **Syllables (earliest use 2-3 months) - for the model**
    - Phonemes (earliest use ~9 months?)
    - Lexical Stress Patterns (earliest use ~8 months)
- Model Assumptions/Concessions:
    - Adult Syllabification & Maximum-Onset Principle vs ???
    - Phoneme-based model vs Phones
    - Syllabification occurs within words

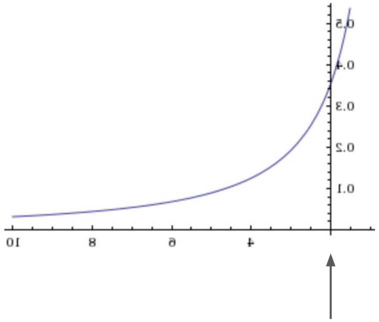


Output

- Useful Oversegmentations - The “All,” “Right,” ~~“Alright,”~~ “-ly,” “-ing,” lexicon
    - Lead to better segmentation elsewhere because they serve as markers - you get the beginning and the end of two other words every time they appear
- Useful Undersegmentations - The “couldi” “isthata” “lookatthekitty” lexicon
    - Help produce syntactic rules
- In the end, we fudge things in the models favor as long as it fits a “useful” pattern
    - In this sense, the paper is very exploratory

Inference

- Ideal “BatchOpt” (MCMC algorithm)
- Incremental Processing “OnlineOpt”
- Sub-optimal Decision Making “OnlineSubOpt” (Probability Matching)
- Recency Effect “OnlineMem” (Decayed MCMC)
  - Probability of sampling a boundary proportional to b_a^{-d}
    - b_a is the number of boundaries until the end of the current utterance
    - d is the decay rate



Current Boundary

Model Evaluations

$$Precision = \frac{\#correct}{\#guessed} = \frac{\#true\ postives}{\#true\ positives + \#false\ positives}$$

$$Recall = \frac{\#correct}{\#true} = \frac{\#true\ positives}{\#true\ positives + \#false\ negatives}$$

$$F\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall}$$

- Word Tokens used as Unit (*the penguin eats the fish* = 5)
- Separate Training & Test Sets

Model Evaluations

		Unigram		Bigram	
		Phoneme	Syllable	Phoneme	Syllable
Bayesian	BatchOpt	55.0 (1.5)	53.1 (1.3)	69.6 (1.6)	77.1 (1.4)
	OnlineOpt	52.6 (1.5)	58.8 (2.5)	63.2 (1.9)	75.1 (0.9)
	OnlineSubOpt	46.5 (1.5)	63.7 (2.8)	41.0 (1.3)	77.8 (1.5)
	OnlineMem	60.7 (1.2)	55.1 (0.3)	71.8 (1.6)	86.3 (1.2)
Other	Lignos 2012	7.0 (1.2)	87.0 (1.4)		
	TPminima	52.6 (1.0)	13.0 (0.4)		

- When compared to adults

Model Evaluations

	Unigram			Bigram		
	Real	Morph	Func	Real	Morph	Func
BatchOpt	0.73	0.13	4.40	4.19	0.69	6.37
OnlineOpt	2.15	0.47	3.17	6.44	0.90	4.85
OnlineSubOpt	2.59	0.45	3.38	8.77	2.08	2.87
OnlineMem	2.19	0.31	5.02	14.41	3.20	3.64
Lignos 2012	19.00	3.59	0.03			
TPminima	0.01	0.00	7.33			

- “Real” and “Morph” represent Oversegmentations
- “Func” represents Undersegmentations

Modeling Human Performance - Frank et al., 2010

- Facets of the data that ought to be more challenging for a human to “solve” word segmentation make it more difficult for Bayesian, but not other, models to do just that.