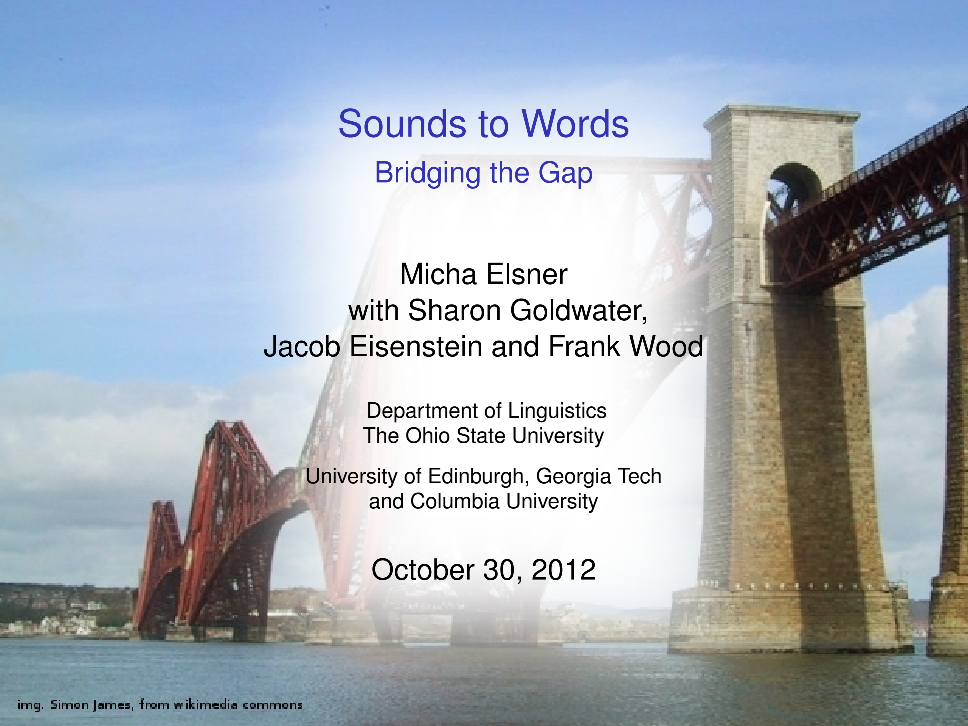


Sounds to Words

Bridging the Gap

Micha Elsner
with Sharon Goldwater,
Jacob Eisenstein and Frank Wood

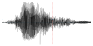
Department of Linguistics
The Ohio State University

University of Edinburgh, Georgia Tech
and Columbia University

October 30, 2012

img. Simon James, from wikimedia commons

Early language learning



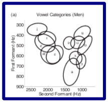
Audio

Interpretable

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/, myhousecallmd.com

Early language learning

Phonetic categories (audio : phoneme)

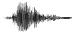


Pronunciation dictionary (phonemes : word)

you *ju, jə, jɪ*
want *wan, want*
...

Language model (words : sentence)

you want
you like
...



Audio

jəwanəkuki

Phonetic transcript

you
want
a
cookie

Segmented

Interpretable

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/

Synergy

- Are /u/ and /i/ different vowels?
  - Pronunciation: yes, because /ju/ is common and /ji/ is rare

Synergy

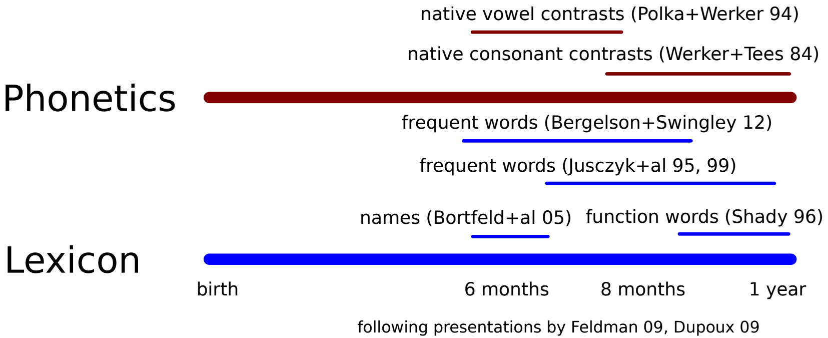
- Are /u/ and /i/ different vowels?
  - Pronunciation: yes, because /ju/ is common and /ji/ is rare
- Is /ju/ a word?
  - Phonetics: yes, because /j/ predicts /u/-like vowel tokens
  - Language model: yes, because it helps predict subsequent /want/

Synergy

- Are /u/ and /i/ different vowels?
  - Pronunciation: yes, because /ju/ is common and /ji/ is rare
- Is /ju/ a word?
  - Phonetics: yes, because /j/ predicts /u/-like vowel tokens
  - Language model: yes, because it helps predict subsequent /want/
- Is /ju/ /want/ different from /ðɛj/ /want/?
  - Pronunciation: yes, because they contain dissimilar segments

Components interact to solve the problem...

Evidence from development



Key developments at roughly the same time

Want to show: these synergies are real!

Cognitive/Linguistic

- Establish role of synergy in early acquisition
- Propose mechanisms: predict developmental stages
- Universals vs. generic learning

Want to show: these synergies are real!

Cognitive/Linguistic

- Establish role of synergy in early acquisition
- Propose mechanisms: predict developmental stages
- Universals vs. generic learning

Applied

- Unsupervised speech recognition
- Learn new lexical items/accents

Related work

- (Martin, Peperkamp, Dupoux ‘12) Clusters symbolic phones into phonemes by learning a proto-lexicon
- (Feldman, Griffiths, Morgan ‘09) Clusters acoustic tokens into phonemes based on a known lexicon
- (Plaut, Kello ‘98) Neural network model of phonetic articulations from known lexicon, uncertain semantics
- (Neubig et al ‘12), (Rytting, Brew 2008) Learn words given uncertain representation of input
- (Vallabha+al ‘07, Varadarajan+al ‘08, Dupoux+al ‘11, Lee+Glass ‘12) Discover phone-like units from acoustics (no lexicon)

Related work

- (Martin, Peperkamp, Dupoux ‘12) Clusters symbolic phones into phonemes by learning a proto-lexicon
- (Feldman, Griffiths, Morgan ‘09) Clusters acoustic tokens into phonemes based on a known lexicon
- (Plaut, Kello ‘98) Neural network model of phonetic articulations from known lexicon, uncertain semantics
- (Neubig et al ‘12), (Rytting, Brew 2008) Learn words given uncertain representation of input
- (Vallabha+al ‘07, Varadarajan+al ‘08, Dupoux+al ‘11, Lee+Glass ‘12) Discover phone-like units from acoustics (no lexicon)

This work

Large, semi-realistic corpus of symbolic input
Learns explicit lexicon and phonetic rules
Future work could integrate some other models!

In this talk

Motivation

Word segmentation: previous work on the lexicon

- Goldwater's Bayesian model of lexical acquisition

Modeling phonetic variation (ACL '12)

- Our Bayesian model
- Channel model: transducer with articulatory features
- Bootstrapping the model
- Greedy inference
- Performance

Jointly segmenting and modeling variation

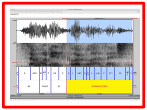
- Inference with beam sampling

Conclusions

Word segmentation

(Setup follows (Brent '99))

Human transcriber (audio : sentence)



Deterministic dictionary (word : phonemes)

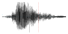
you *ju*
want *want*
…

Deterministic dictionary (phonemes : word)

you *ju*
want *want*
…

Language model (words : sentence)

you want
you like
…



Audio

you want a cookie

Orthographic

juwantəkuki

Normalized phonetic transcript

ju want instead of *jə wan*

you
want
a
cookie

Segmented

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/

The input

Input from phonetic dictionary: why?

- Pipeline model?
  - Learn phonetics first
  - Use learned phonetics to normalize input
- Little theoretical justification for this...
- Real phonetic transcription is *expensive*!
  - Usually requires linguists
  - Very time-consuming
  - Some for adult speech, no child-directed corpora to my knowledge

Mostly a matter of convenience!

Segmenting words: previous work

Previous models use two kinds of evidence:

Boundary-based

*/pɛtkɪti/ : *tk* so */pɛt/ /kɪti/*

- Learn about phonotactics
- Place boundaries to break infrequent sound sequences
- Words defined implicitly by boundary position

(Fleck ‘08, Rytting ‘07, Daland+al ‘10, others)

Lexical

/pɛtkɪti/ : kɪti probably a word, so */pɛt/ /kɪti/*

- Learn probable lexical items
- Propose word sequence to cover observed corpus
- Boundaries defined implicitly by word sequence

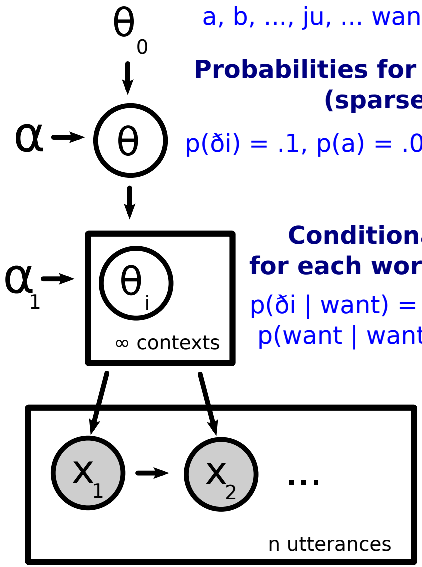
(Brent ‘99, Venkataraman ‘01, Goldwater ‘09, others)

Goldwater et al ‘09

A lexical model of word segmentation:

- Generative Bayesian model
- Two parts: probability of lexicon
  - Dirichlet process: allows infinite, favors small
- Probability of corpus
  - Rewards predictability
- Basis for other work in this talk

Generative story



Infinite list of possible words

a, b, ..., ju, ... want, ... juwant, ...

Probabilities for each word (sparse)

p(ði) = .1, p(a) = .05, p(want) = .01...

Conditional probabilities for each word after each word

p(ði | want) = .3, p(a | want) = .1,
p(want | want) = .0001...

Observed corpus

ju want ə kuki
ju want ɪt
...

What's going on?

Memorizing the data

Lexicon: *juwantəkuki, juwantıt*
Likelihood of corpus is high...
But lexicon is huge: sparse prior says not very likely

Character by character

Lexicon: *j, u, w...*
Lexicon is very sparse: prior is high
Likelihood of corpus is poor

True lexicon

Lexicon: *ju, want...*
A "happy medium"

Goldwater suffers under variation

Goldwater run on Buckeye corpus (Fleck '08)

- Must represent each pronunciation separately
- No var. (dictionary) versus phonetic transcript

	Break F	Token F	Lexicon (type) F
no var.	84	68	27
transcribed	65	35	13

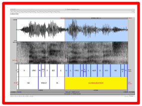
Break F declines: huge decrease in precision

- Undersegmentation

Overview

(Simple) phonetic variation

Human transcriber (audio : sentence)



Pronunciation dictionary (word : phonemes)

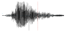
you *ju, jə, jɪ*
want *wan, want*
...

Pronunciation dictionary (phonemes : word)

you *ju, jə, jɪ*
want *wan, want*
...

Language model (words : sentence)

you want
you like
...



Audio

you want a cookie

Orthographic

jə
wan
ə
kuki

Segmented phonetics

you
want
a
cookie

Segmented

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/

Noisy channel setup

loosely based on Goldwater+al 09

random lexicon

want, ju...

word-to-word transition probabilities

p(want|ju), p(to|want)

intended utterances

ju want wan
want e kʊki

our contribution

noisy channel

character sequence rewrite probabilities

p(u → ə : j_$)

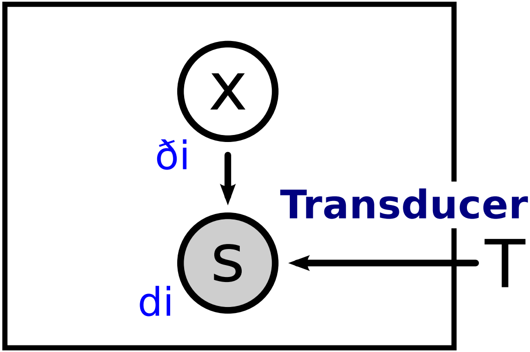
surface (observed)

jə wa? wʌn
wan ə kʊki

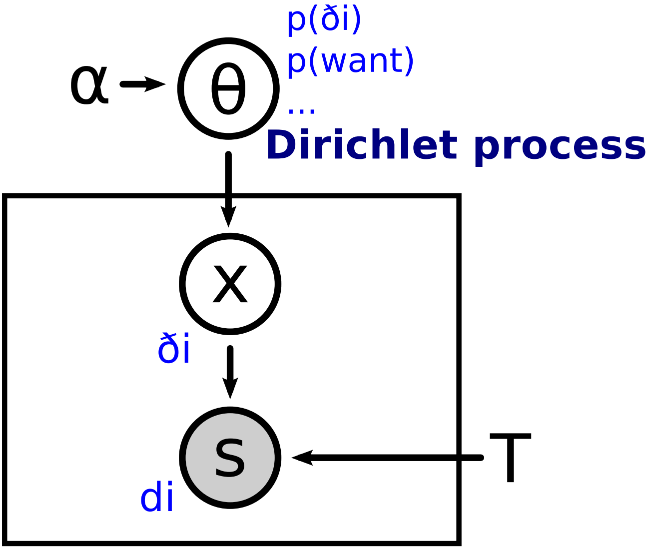
Graphical model

Presented as Bayesian model to emphasize similarities with (Goldwater+al '09)

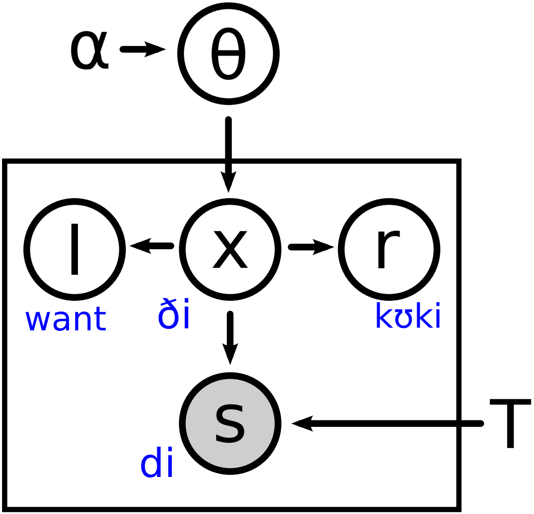
- Our inference method approximate



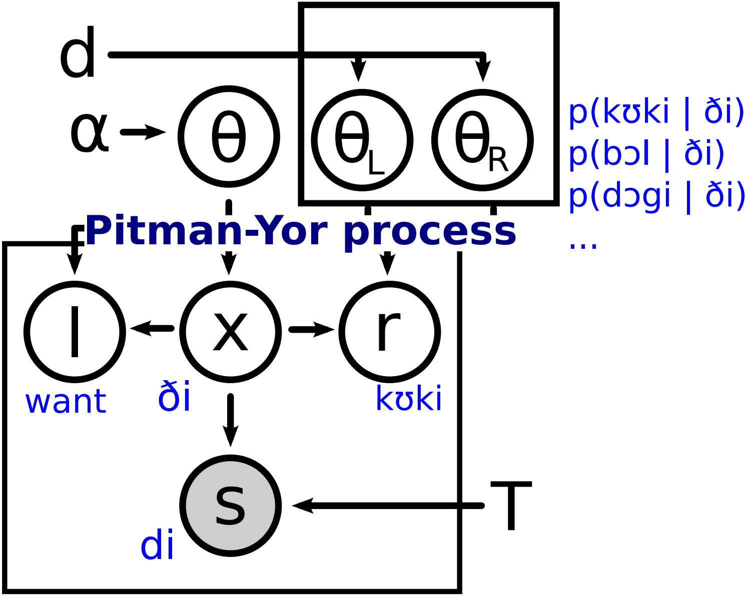
Graphical model



Graphical model



Graphical model



Factorization

Minor point: here we factor:

$$p(l, x, y) = p(x)p(l|x)p(r|x)$$

This generates words twice if we look at the whole corpus...
In this section we only look at subsets of words.
Later we switch to:

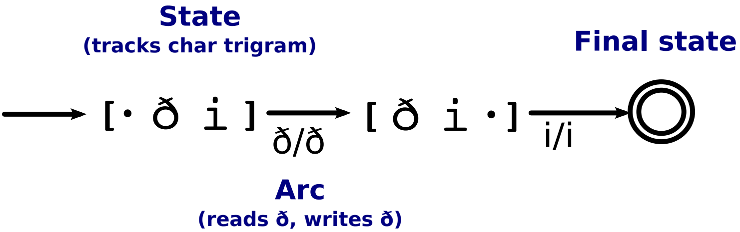
$$p(l, x, y) = p(l)p(x|l)p(r|x)$$

Transducers

Weighted Finite-State Transducer

Reads an input string
Stochastically produces an output string
Distribution $p(out|in)$ is a hidden Markov model

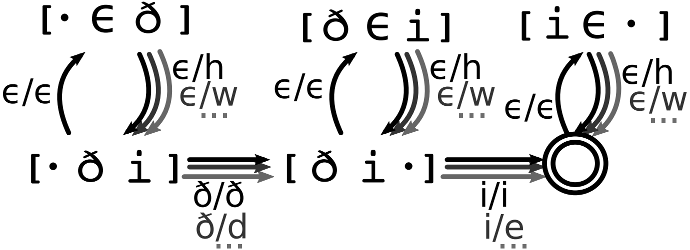
Identity FST given ði
(reads ði "the" and writes ði)



Our transducer

Produces any output given its input
Allows insertions/deletions

Reads ði, writes anything
(Likely outputs depend on parameters)



Probability of an arc

How probable is an arc? $[\cdot\ ð\ i\ ] \xrightarrow[ð/d]{}$

Log-linear model

Extract features f from state/arc pair...

- Score of arc $\propto exp(w \cdot f)$

following (Dreyer+Eisner '08)

Articulatory features

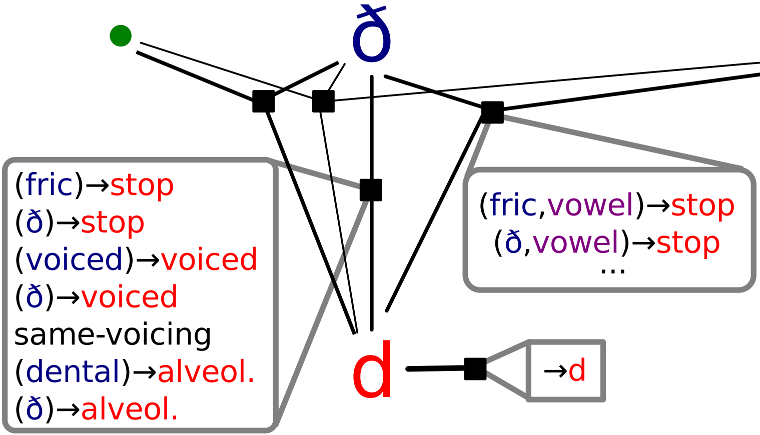
- Represent sounds by how produced
- Similar sounds, similar features
  - ð: voiced dental fricative
  - d: voiced alveolar stop

see comp. optimality theory systems (Hayes+Wilson '08)

Feature templates for state (prev, curr, next) $\rightarrow$ output

Templates for voice, place and manner

Ex. template instantiations:



Learned probabilities

- ð i $\rightarrow$

ð	.7
n	.13
θ	.04
d	.02
z	.02
s	.01
ϵ	.01
...	...

Inference

Bootstrapping

Initialize: surface type $\rightarrow$ itself ([di] $\rightarrow$ [di])

Alternate:

- Greedily merge pairs of word types
  - ex. intended form for all [di] $\rightarrow$ [ði]
- Reestimate transducer

Inference

Bootstrapping

Initialize: surface type $\rightarrow$ itself ([di] $\rightarrow$ [di])

Alternate:

- Greedily merge pairs of word types
  - ex. intended form for all [di] $\rightarrow$ [ði]
- Reestimate transducer

Greedy merging step

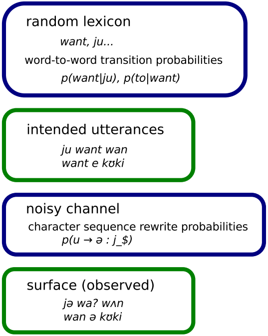
Relies on a **score** Δ for each pair:

- $\Delta(u, v)$: approximate change in model posterior probability from merging $u \rightarrow v$
- Merge pairs in approximate order of Δ

Computing Δ

$\Delta(u, v)$: approximate change in model posterior probability from merging $u \to v$

- Terms from language model
  - Encourage merging frequent words
  - Discourage merging if contexts differ
  - See the paper
- Terms from transducer
  - Compute with standard algorithms
  - (Dynamic programming)



Dataset

We want: child-directed speech,
close phonetic transcription

Use: Bernstein-Ratner (child-directed)

(Bernstein-Ratner '87)

Buckeye (closely transcribed) (Pitt+al '07)

Sample pronunciation for each BR word from Buckeye:

- No coarticulation between words

"about"

ahbawt:15, bawt:9, ihbawt:4, ahbawd:4, ihbawd:4, ahbaat:2, baw:1, ahbaht:1, erbawd:1, bawd:1, ahbaad:1, ahpaat:1, bah:1, baht:1

Evaluation

Map system's proposed intended forms to truth

- {ði, di, ðə} cluster can be identified by any of these
- System doesn't do "phonology"— at this stage, neither may infant?

- Score by tokens; emphasis on frequent words
- ...and types (lexicon); all lexemes counted equally

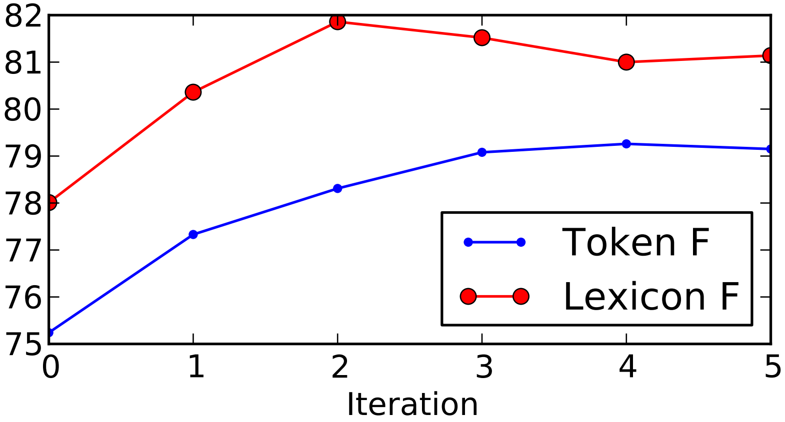
With gold segment boundaries

Scores (correct forms)

	Token F	**Lexicon (Type) F**
Baseline (ident)	65	67
Initializer	75	78
No context	75	76
Full system	**79**	**87**
Upper bound	91	97

Learning

Initialized with weights on *same-sound*, *same-voice*, *same-place*, *same-manner*



Induced word boundaries

Induce word boundaries with (Goldwater+al '09)
Cluster with our system

Scores (correct boundaries and forms)

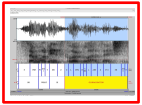
	Token F	Lexicon (Type) F
Baseline (ident)	44	43
Full system	**49**	**46**

After clustering, remove boundaries and resegment: no improvement
Suggests joint segmentation/clustering

Overview

Joint segmentation and word forms

Human transcriber (audio : sentence)



Pronunciation dictionary (word : phonemes)

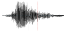
you	*ju, jə, jɪ*
want	*wan, want*
...	

Pronunciation dictionary (phonemes : word)

you	*ju, jə, jɪ*
want	*wan, want*
...	

Language model (words : sentence)

you want
you like
...



Audio

you want a cookie

Orthographic

jəwanəkuki

Phonetic transcript

you
want
a
cookie

Segmented

details Feldman et al 09, http://www.contrib.andrew.cmu.edu/, http://blogs.oucs.ox.ac.uk/

Challenges

Model from previous section fine for joint segmentation/clustering

- (Factor $p(l)p(x|l)p(r|x)$ but this is trivial fix)

Issue is inference:

Challenges

Model from previous section fine for joint segmentation/clustering

- (Factor $p(l)p(x|l)p(r|x)$ but this is trivial fix)

Issue is inference:

- Standard segmentation: sample locations of boundaries
- Only two steps to slice out *want*
- *juwanttu*, **ju**•*wanttu*, **ju•want•tu**

Challenges

Model from previous section fine for joint segmentation/clustering

- (Factor $p(l)p(x|l)p(r|x)$ but this is trivial fix)

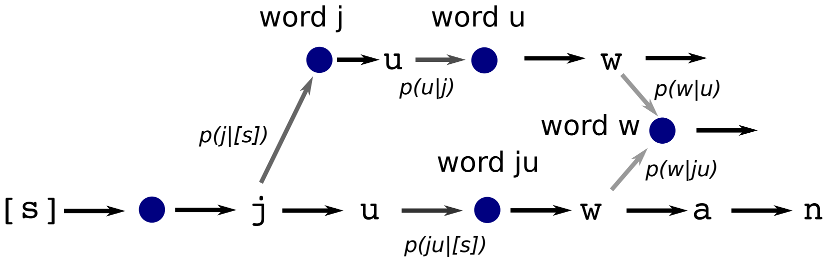
Issue is inference:

- Standard segmentation: sample locations of boundaries
- Only two steps to slice out *want*
- *juwanttu*, **ju**•*wanttu*, **ju•want•tu**

- With clustering, have farther to travel
- *jəwantu*, *jə•wantu*, *jə•wan*•**tu**, **ju**•*wan*•**tu**, **ju•want•tu**
- Need moves that alter multiple letters/boundaries at once

Markov-style sampling methods

Can write a Goldwater model as a big FSM:

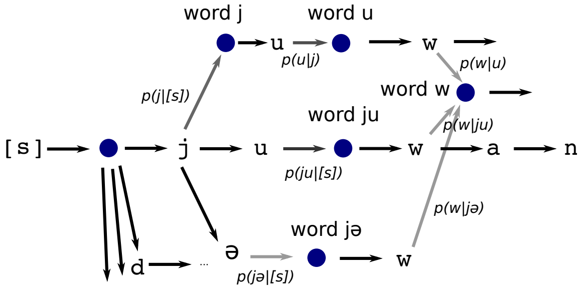


(first in (Mochihashi+al '09))

Only need states that generate the original string

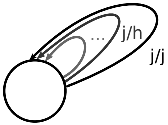
Use forward-backward (plus MH) for inference

Composing with transducer



$\circ$

Unigram transducer
(read any, write any)

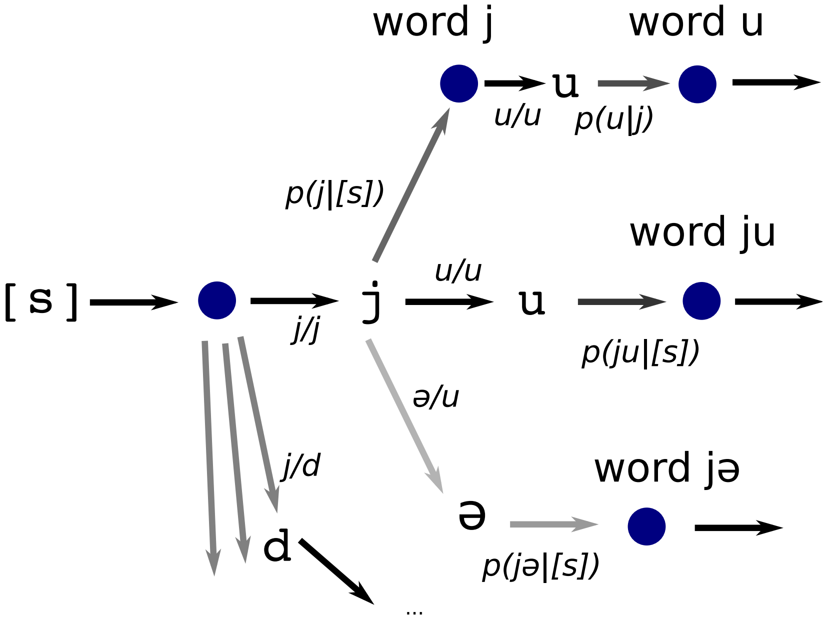


Acceptor for original string

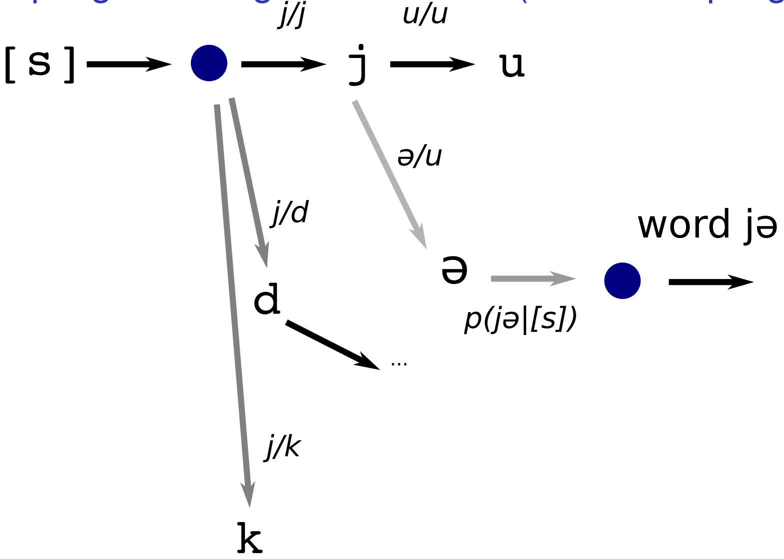
$\circ$

[s] → j → u → w → a → n

Result is a REALLY BIG transducer

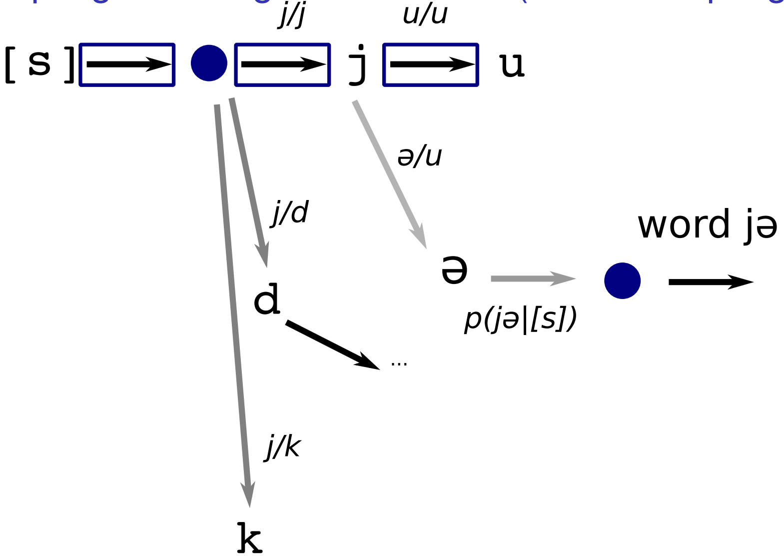


Sampling from huge transducers (beam sampling)



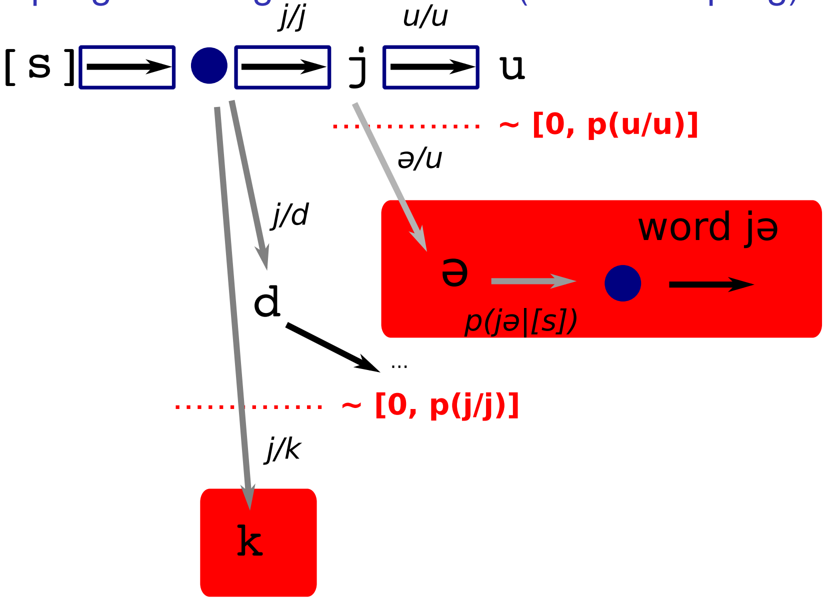
(van Gael+al '08), (Huggins+Wood '12)

Sampling from huge transducers (beam sampling)



(van Gael+al '08), (Huggins+Wood '12)

Sampling from huge transducers (beam sampling)



(van Gael+al '08), (Huggins+Wood '12)

Making this work in practice

Different cutoffs

- Separate cutoffs for letter and word transitions
- Letter cutoffs critical in discarding bad hypotheses
- Can't be *too* different: introduces bias!

The infinite prior

- Prior over words is infinite: so is FST!
- Original paper uses sampling to deal with this: not efficient enough
- Treat prior as another FST...
  - But this introduces bias as well!
- Need to use Metropolis-Hastings rejection step (but usually accept)

Search strategies

Changing one utterance at a time does not collapse common variants:

wʌt, *wʌd*
ju, *jə*

Too many steps needed to convert all tokens...

Search strategies

Changing one utterance at a time does not collapse common variants:

wʌt, *wʌd*
ju, *jə*

Too many steps needed to convert all tokens...

Phase of maximizing word sequence probabilities

- Using two different annealing rates
- Rates $>>$ 1 for word sequence maximize LM probs
- Overgeneralizes lexical items...
- Bad mergers usually unmerge when phase ends

Developmental speculation

System temporarily overgeneralizes words

- Group ðis, ðat, ðey
- Or hypothesize inserted/deleted segments: ɛn and ɛni^j
- Short, vowel-heavy words particularly vulnerable

Evidence from development?

- Don't know any proposals of this theory
- (Merriman+Schuster '99): 2-4 year olds think "japple" might mean "apple" under some circumstances
- Tomasello and others: children learn multiword "chunks"
- Can these be reinterpreted as evidence for phonetic overgeneralization?

Perhaps can test via new experiments...

Preliminary experiment

1000 line dataset

Tokens (boundaries only)	P	R	F
No channel	56	69	62
Joint	64	69	66

Tokens (bounds and forms)	P	R	F
No channel	40	50	45
Joint	50	54	52

Initial finding

Model with channel is better segmenter

- Better precision, fewer breaks overall

Much better at predicting intended forms

- Reassuring but not really surprising

Conclusions

- Data with variations is problematic for models of early lexical acquisition
- Possible to learn phonetics jointly with LM
- Learning synergy improves performance
- Seems possible to do everything jointly...
  - But requires some constraints in learning

Implications and future work

Getting the rest of the way to acoustics will be tricky

- Perhaps fully joint model like (Feldman+al '09)?
- Or pre-clustering like (Varadarajan+al '08)?

Probably some hidden surprises... results here show variation can be very problematic!

Implications and future work

Getting the rest of the way to acoustics will be tricky

- Perhaps fully joint model like (Feldman+al '09)?
- Or pre-clustering like (Varadarajan+al '08)?

Probably some hidden surprises... results here show variation can be very problematic!

Mechanisms for inference require some constraints

- The number of hypotheses our learner considers is vast...
- Keeping it manageable requires multiple interacting random filters

More study needed to find what infants are doing

Thanks

Mary Beckman, Laura Wagner and Lacqueys; Eric Fosler-Lussier, William Schuler and Clippers; funded by EPSRC; thanks for listening!